HAL open science

Identification des reprises et interprétation automatique des expressions pronominales dans des textes en français

François Trouilleux

▶ To cite this version:

François Trouilleux. Identification des reprises et interprétation automatique des expressions pronominales dans des textes en français. Linguistique. Université Blaise-Pascal, Clermont-Ferrand, 2001. Français. NNT : . tel-01152394

HAL Id: tel-01152394
https://hal.science/tel-01152394

Submitted on 16 May 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License

François TROUILLEUX

Identification des reprises et interprétation automatique des expressions pronominales dans des textes en français

Nouveau Doctorat d'Université en Sciences du Langage Linguistique et Informatique

Présenté et soutenu le 21 décembre 2001
devant le jury composé de :

M. Gabriel G. BÈS	Directeur
M. Cassian BRACONNIER	Examinateur
M. Francis CORBLIN	Rapporteur
M. Marcel CORI	Rapporteur
M. Éric GAUSSIER	Examinateur
Mme. Annie ZAENEN	Examinatrice

Laboratoire de Recherche sur le Langage
Université Blaise Pascal - Clermont-Ferrand
GRIL
UFR Langues Appliquées et Communication

Remerciements

Pour tout ce qu'il m'a enseigné, pour son exigence de rigueur, pour sa gentillesse, pour la présente thèse, qui sans lui n'aurait pas vu le jour, j'adresse à Gabriel G. Bès mes remerciements les plus chaleureux. De mon inscription en DEA jusqu'à la fin de la thèse, il a suivi et dirigé mon travail avec un intérêt constant. Je lui suis reconnaissant de m'avoir guidé dans la voie de la démarche scientifique en linguistique et espère avoir été à la hauteur de la confiance qu'il m'a témoignée. Puisse l'avenir nous réserver de nouvelles collaborations.

Je remercie Annie Zaenen et Éric Gaussier, pour leur suivi attentif de mon travail au Centre de recherche de Xerox. Leurs commentaires sur les diverses versions préliminaires de la thèse m'ont été précieux. Sur bien des points, en particulier ce qui concerne les procédures et mesures d'évaluation, l'aide qu'Éric Gaussier m'a apportée a été décisive. À cet égard, les chapitres 3 et 4 de la thèse décrivent un travail que nous avons vraiment mené en commun.

Je remercie les membres du jury de soutenance, Cassian Braconnier, Francis Corblin, Marcel Cori, Éric Gaussier et Annie Zaenen, d'avoir accepté d'exprimer leur jugement sur mon travail. Je remercie plus particulièrement les rapporteurs, Francis Corblin et Marcel Cori, dont les commentaires et suggestions m'ont permis d'améliorer le contenu de la thèse.

Ces trois années furent remplies de longues discussions, enrichissantes pour moi et pour la thèse, avec Salah Aït-Mokhtar, Jean-Pierre Chanod, Marie-Hélène Corréard, Annette Frank, Caroline Hagège, Aarne Ranta et Claude Roux. Je dois également des remerciements plus particuliers à Caroline Hagège pour sa lecture critique et constructive de la thèse et à Claude Roux pour sa patience et sa réactivité face à mes questions concernant le système XIP.

Merci à Pierre Isabelle et Ágnes Sandor pour leurs commentaires sur divers chapitres, ainsi qu'à Catherine Clouzot et Agnès Tutin pour nos fructueux échanges de vues dans notre projet d'annotation de corpus.

Merci à Faïza Abbaci pour son travail d'implantation d'un programme qui traduit au format XML les fichiers annotés dans le système d'annotation décrit dans la première partie de la thèse.

Enfin, je tiens à exprimer mon amitié pour les membres du GRIL et du Centre de recherche de Xerox, présents ou passés, qui ont constitué une excellente compagnie. Parmi eux et outre les personnes déjà citées, je tiens à mentionner plus particulièrement les noms de Mohammed Abaidi, Alain Couillault, Gyöngyvér Forintos-Kosten, Thomas Pfuhl, Emmanuelle Rodier et José Rodrigo, pour le GRIL, et ceux de Christopher Brewster, Caroline Brun, Marc Dymetman, Nuria Gala-Pavia, Bernard Jacquemin, Veronika Lux, Sylvain Pogodalla, Frédérique Segond et Jean-Yves Vion-Dury, pour XRCE.

Je dédie ce travail à Marie-Elisabeth Leca, Joseph, Judith et Suzanne.

Sommaire

Introduction **13**

I Identification des reprises **19**

1 Délimitation du champ d'observation **21**
1.1 Anaphore 22
1.2 Reprises 25
1.2.1 Liens caractérisés par ou sans recours à l'identité 26
1.2.2 Liens de reprise non anaphoriques 28
1.2.3 Précisions 29
1.3 Motivations 30
1.3.1 Une approche centrée sur le résultat de l'interprétation . . 30
1.3.2 Des observations faisant l'objet d'une inter-subjectivité . . 31
1.4 Terminologie 32
1.4.1 Univers de dénotation 33
1.4.2 Expressions dénotantes 34
1.4.3 Descriptions et expressions non dénotantes 36
1.4.4 Êtres singuliers et ensembles d'êtres singuliers dans l'univers de dénotation 37
1.5 Notation des observations 39
1.5.1 Délimitation des expressions 39
1.5.2 Description du lien observé 40

2 Une typologie des liens de reprises **41**
2.1 Identité de dénotation 42
2.1.1 Définition 42
2.1.2 Cas général 43
2.1.3 Désignation d'ensembles 45
2.2 Identité de description 49
2.2.1 Reprise par une expression dénotante 50
2.2.2 Reprise par une expression non dénotante 52

2.3 Relation « membre-de » 56
2.3.1 Exemples 57
2.3.2 Définition 58
2.3.3 Transitivité de la relation membre-de 60
2.3.4 Relation « membre-de » et identité de description 61
2.3.5 Relation membre-de et identité de dénotation 63
2.4 Relation « distingué-de » 64
2.4.1 Exemples 65
2.4.2 Définition 66
2.4.3 Absence de source explicite pour une reprise avec *autre*. . 68
2.4.4 Relation « distingué-de » et identité de description 68
2.5 Référence à un discours 70
2.6 Au-delà des reprises : anaphore associative et deixis 71
2.6.1 Relations référentielles 72
2.6.2 Référence à une date 74
2.7 Mise en perspective 77
2.7.1 Sémantique formelle 77
2.7.2 La cohésion selon Halliday et Hasan 83
2.7.3 Le schéma MATE d'annotation de la coréférence 85
2.7.4 L'annotation (co-)référentielle selon Salmon-Alt 87

3 Critères d'évaluation 93
3.1 Problématique 94
3.1.1 Intérêt des critères d'évaluation 94
3.1.2 Données de l'évaluation 95
3.2 Évaluer l'assignation des dénotations 97
3.3 Correspondance entre clé et réponse 100
3.3.1 Correspondance de type 1-1 100
3.3.2 Spécificité des descriptions 101
3.4 Calcul de la correspondance 103
3.4.1 Similarité entre deux chaînes de références 104
3.4.2 Correspondance entre chaînes de référence 107
3.5 Exemple 108
3.5.1 Une réponse fictive 108
3.5.2 Correspondance entre la clé et la réponse 108
3.5.3 Deux ensembles de prédicats d'observation 110
3.6 Mesures d'évaluation 111
3.6.1 Rappel et précision 111
3.6.2 Substitution, sur-génération, sous-génération 113
3.7 Mise en perspective 115
3.7.1 Exemples de situations d'évaluation 115
3.7.2 L'approche par liens de Vilain et al. 116

3.7.3 La méthode « classes noyaux exclusifs » de Popescu-Belis 120
3.7.4 L'algorithme B-3 de Bagga et Baldwin 124
3.8 Conclusion 125

4 Test d'opérationalité 127
4.1 Données de l'expérience 128
4.1.1 Trois textes, un expert et cinq annotateurs 128
4.1.2 Instructions aux annotateurs 130
4.1.3 Relations observées par l'expert 130
4.1.4 Organisation des données 131
4.1.5 Évaluation des annotations 131
4.1.6 Documents en annexe 135
4.1.7 Difficultés 136
4.2 Mesures d'évaluation globales 137
4.2.1 Moyenne 138
4.2.2 Variance 139
4.2.3 Opinion majoritaire 143
4.2.4 Seuil d'opérationalité 145
4.3 Vue globale des résultats 146
4.4 Identité de description 147
4.5 Reprises de type paraphrase 148
4.6 Identité de dénotation 150
4.6.1 Vue globale 150
4.6.2 Résultats par types d'expressions 150
4.7 Expressions temporelles 155
4.8 Relations référentielles 156
4.9 Discussion 158
4.9.1 Notions opérationnelles 158
4.9.2 Pistes pour une amélioration des résultats 159
4.10 Conclusion de la première partie 164

II Interprétation automatique des expressions pronominales 167

5 Introduction 169
5.1 Objectif 169
5.1.1 Expressions anaphoriques visées 170
5.1.2 Spécifier l'interprétation des expressions pronominales . . 172
5.1.3 Données spécifiées par la clé 175
5.1.4 Prédicat d'évaluation global 177
5.2 Environnement de travail 178
5.2.1 Analyse morphologique 179

5.2.2 Analyse syntaxique 181
5.2.3 Système de résolution des pronoms 181
5.3 Méthodologie 183
5.3.1 Un système implanté en machine 183
5.3.2 Validité statistique de notre hypothèse 183
5.3.3 Étude de corpus 184
5.4 Plan de la deuxième partie 184

6 Approches du problème 187
6.1 Syntaxe 187
6.1.1 La théorie du liage 188
6.1.2 Contraintes de non-coréférence 189
6.2 Restrictions de sélection 191
6.2.1 Idée générale 192
6.2.2 Utilisation effective des restrictions de sélection 193
6.3 Pragmatique 196
6.4 Cohérence et structure du discours 197
6.4.1 La théorie du centrage 197
6.4.2 La théorie des veines 203
6.5 Systèmes d'interprétation automatique 207
6.5.1 Popescu-Belis : résolution de la référence en français . . . 208
6.5.2 Hobbs : un « algorithme naïf » utilisant seulement l'information morpho-syntaxique 209
6.5.3 Lappin & Leass 210
6.5.4 Kennedy & Boguraev : une approche « sans analyseur » . 212
6.5.5 Baldwin : le système CogNIAC 212
6.5.6 Mitkov : une approche « robuste et pauvre en connaissance » 214
6.5.7 Ge et al. 217
6.5.8 Autres systèmes 217
6.6 Mise en perspective de notre propre système 219

7 Analyse syntaxique en entrée du système 221
7.1 Arbre syntaxique 222
7.1.1 Nœuds lexicaux 222
7.1.2 Syntagmes et propositions noyau 223
7.1.3 Nœuds « phrase » 226
7.2 Le système de traits dans XIP 228
7.2.1 Déclaration des traits 228
7.2.2 Attributs généraux 229
7.2.3 Conditions sur les traits 229
7.2.4 Modes d'assignation des traits 231
7.3 Catégories utilisées pour l'analyse du français 232

7.3.1 Catégorisation des noms 233
7.3.2 Catégorisation des pronoms 241
7.3.3 Catégorisation des déterminants 244
7.3.4 Nombre, genre et personne 245
7.3.5 Prépositions 247
7.3.6 Formes verbales 247
7.3.7 Propositions 248
7.3.8 Phrases 249
7.3.9 Insertions 250
7.3.10 Autres traits utilisés 253
7.4 Dépendances 253
7.4.1 Des relations entre les nœuds de l'arbre syntaxique 254
7.4.2 Inventaire des relations syntaxiques 254
7.4.3 Traits associés aux relations 262
7.5 Apports personnels 263

8 Formalisme 265
8.1 Structure des règles 265
8.2 Expressions régulières 266
8.2.1 Séquences de nœuds 266
8.2.2 Expressions simples 266
8.2.3 Expressions complexes 267
8.2.4 Enchâssements de nœuds 268
8.3 Instanciation des expressions régulières 271
8.3.1 Variables sur les nœuds de l'arbre syntaxique 271
8.3.2 Plus longue ou plus courte séquence 272
8.4 Conditions 274
8.4.1 Conditions sur les traits associés aux nœuds 274
8.4.2 Conditions sur l'existence ou la non-existence d'une relation 275
8.4.3 Conditions sur la précédence, l'identité ou la non-identité 276
8.5 Conclusions des règles 277
8.5.1 Création d'une relation 277
8.5.2 Effacement d'une relation 278
8.5.3 Création et effacement conjoints 279
8.5.4 Assignation de traits 280

9 Organisation du système de résolution 283
9.1 Données 283
9.1.1 Sortie du système 283
9.1.2 Rappel des données de la clé 285
9.1.3 Rappel du prédicat d'évaluation global 285
9.2 Organisation globale du système 286

9.3 Étapes du processus d'analyse 287
9.3.1 Règles sur les expressions dénotantes 287
9.3.2 Règles sur les zones d'antécédence 289
9.3.3 Contraintes 290
9.3.4 Préférences 292
9.3.5 Transitivité des antécédents vers les sources 293
9.4 Organisation de la suite de la présentation 293
9.4.1 Règles et préférences 294
9.4.2 Description détaillée des règles et préférences 294
9.4.3 Justification des règles et préférences 295

10 Règles 297
10.1 Règles sur les expressions dénotantes 297
10.1.1 Définition générale 298
10.1.2 Exceptions 301
10.1.3 Relation entre chaque expression dénotante et la phrase qui la contient 306
10.2 Règles sur les zones d'antécédence 306
10.2.1 Pronoms clitiques 308
10.2.2 Pronoms disjoints 320
10.2.3 Déterminants possessifs 324
10.3 Contraintes 329
10.3.1 Contraintes d'accord 329
10.3.2 Contraintes relationnelles 331
10.3.3 Contraintes sur les insertions 339
10.4 Une propriété générale des règles 341

11 Préférences 343
11.1 Vue générale 343
11.1.1 Organisation des préférences 344
11.1.2 Résumé de l'information utilisée dans les préférences . . . 345
11.1.3 Formulation succincte des préférences 348
11.1.4 Exemples 349
11.2 Description détaillée 351

12 Évaluation 369
12.1 Données de l'évaluation 370
12.1.1 Données de la clé 370
12.1.2 Données en sortie 372
12.1.3 Mesures d'évaluation 373
12.1.4 Situations d'évaluation 374
12.1.5 Prédicats d'évaluation 375

12.1.6 Antécédents et référents 378
12.2 Évaluation globale 380
12.2.1 Répartition des expressions pronominales 381
12.2.2 Jugements sur la sortie du système 382
12.2.3 Mesures d'évaluation en sortie finale 382
12.2.4 Analyse globale des erreurs 385
12.3 Évaluation des règles 388
12.3.1 Règles sur les expressions dénotantes 388
12.3.2 Règles sur les zones d'antécédence 389
12.3.3 Contraintes 397
12.4 Évaluation des préférences 402
12.4.1 Données 402
12.4.2 Évaluation globale 404
12.4.3 Évaluation analytique 405
12.5 Pertinence de l'évaluation 412
12.6 Conclusion et perspectives 414
12.6.1 Apports de notre travail 414
12.6.2 Perspectives 416

Bibliographie **425**

Annexes **435**

A Données du test d'opérationalité **435**
A.1 Annotation de référence 436
A.2 Annotation réponse 1 441
A.3 Annotation réponse 2 446
A.4 Annotation réponse 3 451
A.5 Annotation réponse 4 456
A.6 Annotation réponse 5 461
A.7 Inventaire des liens observés 467

B Exemple d'analyse syntaxique **477**

Introduction

Le travail présenté ici s'inscrit dans le domaine du « traitement automatique des langues », ou « linguistique computationelle ». Il a été réalisé dans le cadre d'une Convention industrielle de formation par la recherche en entreprise (CIFRE) au Centre de recherche européen de Xerox (XRCE), le laboratoire universitaire associé étant le Groupe de recherche dans les industries de la langue (GRIL) de l'université Blaise-Pascal à Clermont-Ferrand.

Une partie du travail de recherche en linguistique informatique à XRCE vise à aboutir à l'implantation effective de systèmes de traitement automatique des langues qui puissent être intégrés dans des logiciels permettant d'extraire rapidement de bases de données textuelles importantes les informations pertinentes pour un utilisateur. Les tâches spécifiées dans le cadre des *Message Understanding Conferences* (MUC) illustrent cette problématique : par exemple, pour MUC-6 [38], il s'agissait, étant donné un large ensemble de textes, d'enregistrer dans une structure de données appropriée toutes les informations concernant les successions à la tête d'entreprises mentionnées dans ces textes (qui succède à qui, à la tête de quelle entreprise, quand a eu lieu le changement de direction, etc.).

Si on se place à un niveau très général, l'objectif est de construire une sorte d'interpréteur automatique de textes, machine que l'on peut caractériser en reprenant les termes de L. Karttunen [51] : idéalement, il s'agirait de construire...

> ... un automate conçu pour lire un texte dans une langue naturelle donnée [1], l'interpréter, et enregistrer en quelque manière son contenu, par exemple pour être en mesure de répondre à des questions sur ce texte. Pour accomplir cette tâche, la machine devra remplir au minimum les exigences suivantes. Elle devra être en mesure de construire un fichier contenant la liste de toutes les entités, événements, objets, etc... mentionnés dans le texte, et pour chaque entité enregistrer ce qui en est dit [2].

Parmi les problèmes que devra résoudre la machine de Karttunen, nous nous intéresserons à l'identification des liens de coréférence, ou, plus généralement, des

[1] Le français dans notre cas.
[2] Ce texte est cité ici dans la traduction de F. Corblin [23].

liens de « reprise ». L'observation du texte suivant donnera un premier exemple des problèmes que nous aborderons :

(1) M. Philippe Guilhaume, nommé président-directeur général d'A2 et de FR3 par le CSA il y a cinq mois, doit affronter une virulente campagne du pouvoir politique qui cherche à le déstabiliser par tous les moyens. Le PDG se défend en relançant des revendications salariales qui embarrassent l'État. Prises en otage dans ce conflit, les chaînes de télévision publiques risquent d'être les premières victimes de ces sombres manœuvres.

Parmi les entités mentionnées dans ce texte, certaines le sont plusieurs fois et par des expressions différentes. Ainsi les expressions *M. Philippe Guilhaume*, *le* dans le syntagme *le déstabiliser*, *Le PDG* et *se* dans *se défend* désignent une seule et même personne ; les chaînes de télévision publiques et l'ensemble constitué d'A2 et FR3 sont une seule et même chose ; les expressions *le pouvoir politique* et *l'État* désignent, sinon une même entité, deux entités qui sont étroitement liées. On notera aussi que le syntagme *ce conflit* désigne l'ensemble de la situation évoquée dans les phrases qui précèdent ; de même, d'une manière un peu différente, pour le syntagme *ces sombres manœuvres*.

Nous avons ici mis en relation certaines des expressions de ce texte, et pour certaines, comme par exemple le pronom *le* ou le syntagme *le PDG*, qui sont des expressions anaphoriques, cette mise en relation était indispensable à leur interprétation. *A priori* le pronom *le* peut être utilisé pour désigner n'importe quoi qui puisse être décrit par un nom masculin et le syntagme *le PDG* pour désigner n'importe quel PDG. En disant que ces expressions doivent être interprétées comme l'expression *M. Philippe Guilhaume*, nous avons restreint leur interprétation.

Identifier des relations entre les expressions d'un texte, telles que celles que nous avons mises à jour dans ce court texte, est la tâche à laquelle nous nous intéresserons. Notre étude portera donc essentiellement sur l'interprétation des textes, plus précisément sur l'identification de ce que nous appelons les « liens de reprises » entre expressions, notion qui recouvre la notion de « coréférence » et une partie des phénomènes couverts par la notion d'anaphore.

Il faut noter que notre objectif, dans la thèse, ne sera pas de spécifier dans son intégralité une machine telle que celle décrite par Karttunen, mais d'apporter quelques éléments d'information en vue de son implantation et de son évaluation : nous présenterons, dans une première partie, une étude générale des « phénomènes de reprises », et, dans une seconde partie, la spécification et l'implantation d'un système d'interprétation automatique de certaines expressions pronominales.

Il y a une certaine distance entre les phénomènes décrits dans la première partie de la thèse et ceux pour lesquels nous proposerons effectivement un système d'interprétation automatique dans la seconde partie : les expressions pronominales traitées par le système décrit dans la seconde partie constituent en effet

un ensemble réduit des phénomènes de reprise. Définir et implanter effectivement un système d'interprétation pour l'ensemble des phénomènes de reprise dépasse largement le cadre d'une thèse. Néanmoins, une telle limitation n'invalide pas la démarche suivie dans la première partie de la thèse, démarche qui a consisté à proposer une vue générale des phénomènes de reprise, à spécifier les critères d'évaluation pour la tâche d'identification des reprises et à évaluer l'inter-subjectivité des observations sur cette tâche. Cette problématique constitue la première étape d'un travail de longue haleine, qui est entamé dans la seconde partie de la thèse avec l'implantation d'un système d'interprétation automatique de certaines expressions pronominales.

À un niveau plus général, le fil directeur de la thèse peut être vu dans notre volonté d'illustrer une méthode de travail qui exige que les hypothèses que nous serions par la suite susceptibles de formuler sur les mécanismes qui régissent les phénomènes de reprise puissent être et soient évaluées, c'est-à-dire confrontées à la réalité observable des textes. La problématique de cette démarche scientifique fait dans une large mesure l'objet de la première partie de la thèse ; dans la seconde partie, nous mettons effectivement cette démarche en œuvre en présentant un système d'hypothèses explicite, testable, effectivement testé et évalué.

Identification des reprises

La première partie de la thèse a pour double objectif de présenter une étude générale des phénomènes de reprise et d'illustrer une méthodologie, dans une très large mesure celle que défend G. Bès dans le paradigme 5P [10] et qui met l'accent sur la nécessité de tester les hypothèses par rapport au réel observable.

Le premier chapitre est consacré à la délimitation du champ d'observation : les phénomènes de reprise. Il y a reprise lorsqu'entre deux expressions d'un même texte existe un lien sémantique caractérisé en ayant recours à une relation d'identité. La notion de reprise recouvre la relation de coréférence, mais également d'autres relations. Elle recouvre également certains phénomènes d'anaphore, mais, contrairement à cette dernière, elle ne met pas en jeu une caractérisation du phénomène par la forme des expressions. La caractérisation des différents types de liens de reprise fait l'objet du chapitre 2.

Les deux premiers chapitres ont pour vocation de spécifier un système d'organisation des données linguistiques : on définit les objets qui devront être observés, et comment ils devront l'être. Étant donné un texte quelconque et les définitions des chapitres 1 et 2, un observateur quelconque doit pouvoir spécifier quels sont les liens de reprise à l'intérieur de ce texte. Dans le même temps qu'on décrit les données linguistiques, dans le chapitre 2, on se dote d'un système de notation des observations, de manière à disposer d'un langage de description codifié, qui permette la comparaison des observations faites par différents observateurs sur les mêmes données.

Il est nécessaire de pouvoir comparer des observations faites par différents observateurs pour deux raisons. La première a trait à la nécessité d'évaluer les résultats produits par un système d'hypothèses, celui-ci pouvant être vu comme une sorte d'observateur. La seconde a trait à la nécessité d'attester que les conditions d'évaluation du système d'hypothèses en question existent, en montrant que des observateurs différents font bien les mêmes observations sur les mêmes données.

La problématique de l'évaluation est l'objet des chapitres 3 et 4. Le chapitre 3 est plus particulièrement dédié à la définition de critères et de mesures d'évaluation pour les phénomènes relevant de la coréférence au sens strict. Il s'agit de définir les jugements que nous pourrons être amenés à porter sur un ensemble d'observations E_i par comparaison avec un autre ensemble d'observations E_j effectuées sur les mêmes données : qu'est-ce qu'une observation correcte ?, qu'est-ce qu'une observation incorrecte ?, etc., dans les limites du champ d'étude défini. On se dote par ailleurs de mesures d'évaluation pour évaluer le degré d'adéquation de l'ensemble d'observations E_i relativement à l'ensemble E_j. Des critères et mesures d'évaluation pour la coréférence existaient préalablement à cette thèse ; nous en proposons de nouveaux.

Supposons qu'on veuille définir un système d'hypothèses sur les mécanismes qui régissent les phénomènes de reprise, système aboutissant éventuellement à l'implantation d'un programme informatique doté des fonctionnalités décrites par Karttunen. Pour attester l'existence de conditions d'évaluation externes à ce système d'hypothèses, il est nécessaire d'attester que l'observation des phénomènes visés est inter-subjective. Cette problématique fait l'objet du chapitre 4. Sont présentés les résultats d'une expérience visant à évaluer l'inter-subjectivité des observations sur un ensemble de phénomènes plus large que ceux qui sont couverts par notre notion de reprise définie aux chapitres 1 et 2. Cinq étudiants du GRIL ont noté les relations qu'ils observaient entre les expressions apparaissant dans trois articles du journal La Tribune (que ces relations soient caractérisées par recours ou sans recours à une relation d'identité), observations que nous comparons avec les observations que nous-mêmes avions faites sur ces textes. Les résultats montrent une absence d'inter-subjectivité sur un certain nombre de relations, absence qui nous a conduit à circonscrire plus précisément, avec la notion de reprise définie aux chapitres 1 et 2, les phénomènes qu'il est raisonnable d'envisager de traiter automatiquement.

À certains égards, la première partie de la thèse présente un travail *en cours*. Notre démarche a été la suivante : nous avons défini dans un premier temps un ensemble de phénomènes à observer, puis nous avons évalué l'inter-subjectivité des observations sur ces phénomènes (chapitre 4). Les résultats de ce test soulèvent plus de questions qu'ils n'apportent de réponses. En première réponse à ces question, nous avons été amenés à restreindre l'ensemble des phénomènes initialement caractérisés et à affiner leur définition (notion de reprise définie au

chapitre 1 et typologie du chapitre 2). Sur ce dernier ensemble de phénomènes, un nouveau test d'inter-subjectivité devrait être conduit, mais poursuivre le travail dans cette voie nous aurait conduit à exclure de la thèse le travail plus applicatif présenté dans la seconde partie.

Interprétation automatique des expressions pronominales

Si la première partie de la thèse donne une vue générale des phénomènes de reprise, la seconde partie a un caractère plus applicatif : on y décrit l'implantation d'un système d'interprétation automatique de certaines expressions pronominales dans les textes en français, en l'occurrence les pronoms personnels et les déterminants possessifs de troisième personne.

Le chapitre 5 décrit notre objectif pour cette seconde partie de la thèse, l'environnement de travail et la méthodologie adoptée.

Le chapitre 6 présente les approches possibles du problème que nous voulons résoudre. On recense d'abord les différentes sources d'information qui peuvent entrer en jeu dans les mécanismes d'interprétation des expressions pronominales (syntaxe, sémantique, etc.), puis on décrit quelques-uns des principaux systèmes d'interprétation automatique des pronoms, par rapport auxquels nous mettons notre propre système en perspective.

Le chapitre 7 décrit les données sur lesquelles seront exprimées nos hypothèses sur l'interprétation des pronoms. Il s'agit essentiellement d'une représentation de la structure syntaxique des phrases, telle que produite par l'analyseur syntaxique développé au Centre de recherche de Xerox.

Nos hypothèses sur l'interprétation des expressions pronominales retenues ont été implantées dans le formalisme de l'outil XIP. La description de ce formalisme fait l'objet du chapitre 8.

Notre système d'hypothèses sur l'interprétation des expressions pronominales est présenté plus spécifiquement dans les chapitres 9 (organisation globale du système), 10 et 11 (description précise des hypothèses). Il est ensuite évalué dans le chapitre suivant.

Première partie

Identification des reprises

Chapitre 1

Délimitation du champ d'observation

La première partie de la thèse a pour objet la présentation d'une typologie de ce que nous appelons les « liens de reprise », la spécification de critères d'évaluation pour la tâche d'identification de tels liens dans les textes et l'évaluation de l'inter-subjectivité de l'observation de ces phénomènes. Nous délimitons dans ce premier chapitre le champ d'observation.

La notion de reprise, telle que nous la concevons, décrit des liens qui peuvent être observés entre des expressions d'un même texte et qui sont caractérisés en ayant recours à une relation d'identité. Elle recouvre en particulier les phénomènes de « coréférence », terme entendu ici dans le sens strict de « lien entre deux expressions qui désignent le même objet » [1], et certains des phénomènes couverts par la notion d'anaphore.

Nous abordons la définition de notre champ d'observation par une présentation préalable de la notion d'anaphore (section 1.1), avec laquelle nous supposons le lecteur familier. Les phénomènes de reprise sont ensuite caractérisés en les mettant en perspective avec la notion d'anaphore (section 1.2) :

- d'une part, parmi les liens anaphoriques qui peuvent être observés, certains, caractérisés par recours à une relation d'identité, seront des phénomènes de reprise, d'autres, caractérisés sans recours à une relation d'identité, non ;
- d'autre part, nous observerons que les liens caractérisés par recours à une relation d'identité sont aussi observables dans des contextes qui ne relèvent pas de l'anaphore.

Nous justifions ensuite notre approche dans la section 1.3 : les idées essentielles sont (i) de faire abstraction de ce qui dans la notion d'anaphore tend à décrire un phénomène qui relève du processus d'interprétation des expressions pour aboutir

[1] C'est le sens donné à ce terme dans les dernières campagnes d'évaluation MUC [44].

à une notion de reprise qui met l'accent sur le seul résultat de l'interprétation et (ii) de nous concentrer sur des phénomènes pour lesquels les observations ont le plus de chance d'être inter-subjectives.

Enfin nous introduisons la terminologie qui sera utilisée dans la suite de la thèse (section 1.4) ainsi que la base d'un système de notation des observations (section 1.5).

1.1 Anaphore

La notion d'anaphore, dans son acception la plus courante en linguistique, décrit une relation entre deux expressions d'un même texte, relation telle que l'interprétation de l'une des deux expressions, dite « expression anaphorique », est spécifiée par l'interprétation de l'autre, que nous appellerons la « source » [2] de l'expression anaphorique. Les exemples qui suivent illustrent quelques phénomènes d'anaphore. Les expressions anaphoriques sont notées en italiques et leur source en petites capitales [3].

(1) JACQUES dort. *Il* est fatigué.

(2) FORTIS décroche enfin le contrôle de LA GÉNÉRALE DE BANQUE. *Les deux établissements* vont former le premier groupe bancaire en Belgique.

(3) Marie aime LES PLAGES DE L'ATLANTIQUE ; Pierre préfère *celles de la Méditerranée*.

(4) Marie aime LES FLEURS BLANCHES ; Pierre préfère *les rouges*.

(5) Parmi CES QUATRE ADMINISTRATEURS, *trois* ont démissionné.

(6) Ne placez pas L'IMPRIMANTE à un endroit où des personnes pourraient marcher sur *le câble d'alimentation*.

(7) Nous entrâmes dans UN VILLAGE ; *l'église* était sur une hauteur.

(8) Une semaine après le TREMBLEMENT DE TERRE qui a violemment secoué, le 17 octobre, la région de San-Francisco, le nombre *des victimes* s'établit à 63 morts et 9 disparus.

On notera que dans certains cas, l'interprétation d'une expression anaphorique peut être spécifiée non pas par *une* expression, mais par plusieurs expressions. C'est le cas dans l'exemple (2), où l'expression anaphorique *les deux établissements* désigne l'ensemble constitué des deux sociétés désignées par ailleurs par les expressions *Fortis* et *la Générale de Banque*.

Précisons également que la notion d'anaphore est utilisée pour décrire la relation entre deux expressions qui ne dépendent pas syntaxiquement l'une de l'autre.

[2] Nous n'employons pas ici le terme couramment employé d'« antécédent », car nous lui donnerons une acception différente de « source » dans la seconde partie de la thèse.

[3] Nous adoptons ici volontairement une délimitation parfois assez large des expressions anaphoriques, par exemple dans l'exemple (3) une alternative aurait été de considérer comme source le seul mot *plages* et comme expression anaphorique le seul mot *celles*.

Dans l'exemple (6), le rapport entre les expressions *le câble d'alimentation* et *l'imprimante* est le suivant : l'interprétation de l'expression *l'imprimante* permet de spécifier l'interprétation de l'expression *le câble d'alimentation* dans le sens où cette dernière expression pourrait être utilisée pour désigner n'importe quel câble d'alimentation, mais désigne en fait le câble d'alimentation de l'imprimante dont on vient de parler. Le même type de remarque pourrait être fait pour les mêmes expressions dans la phrase suivante :

(9) Ne marchez pas sur le câble d'alimentation de l'imprimante.

La situation illustrée en (9) n'est cependant pas jugé comme un cas d'anaphore car le lien entre les deux expressions est explicite dans la structure de la phrase, ce qui n'est pas le cas dans l'exemple (6).

Au-delà de l'idée, exposée dans notre définition initiale, qu'il y a spécification de l'interprétation d'une expression par une autre, la notion d'anaphore est souvent caractérisée relativement à un ensemble particulier d'expressions susceptibles d'être anaphoriques.

F. Corblin, par exemple, met en avant l'idée que c'est la forme de l'expression anaphorique qui « déclenche » l'anaphore [23, p. 41] :

> Même si elle ne s'exprime pas toujours clairement dans les mêmes termes, l'idée qui réunit la plupart des approches de l'anaphore est qu'on a affaire à une opération déclenchée par une forme insuffisamment spécifiée, incomplète ; l'opération de mise en rapport au contexte a pour effet de saturer une forme qui exige de l'être. Il y a de cela quelques témoignages assez nets : il est typique que l'étude de l'anaphore ne peut contourner la question des positions (ou catégories) vides, qu'elle accorde une place privilégiée aux formes peu spécifiées (pronoms) [...]

D'une manière analogue, S. Botley et T. McEnery signalent que l'anaphore se réalise à travers des « marqueurs linguistiques » [12, p. 2] :

> L'anaphore permet à un locuteur de rappeler à la conscience de son interlocuteur les entités ou concepts qui ont déjà été introduits dans le discours. En anglais, par exemple, l'anaphore peut être réalisée par de nombreux marqueurs linguistiques différents, tels que les pronoms, les pronoms démonstratifs, les substitutions pronominales ou les ellipses.

La notion d'anaphore est donc souvent liée à l'emploi de certaines expressions. Les catégories d'expressions le plus souvent retenues comme pouvant être des expressions anaphoriques sont les suivantes [4] :

- les expressions pronominales de troisième personne (les expressions de première et deuxième personne étant considérées comme « déictiques », c'est-

[4] Dans la suite de la thèse, nous limiterons l'usage du terme « expression anaphorique » aux expressions appartenant à l'une des catégories mentionnées ici et dont interprétation est spécifiée par mise en relation avec une autre expression.

à-dire comme étant interprétées relativement à la situation d'énonciation du discours) ;
- les phénomènes d'ellipse [5] (les « catégories vides » de Corblin, p. ex. *Jacques mange une pomme et Juliette* ∅ *une poire.*) ;
- les syntagmes nominaux démonstratifs (p. ex. *ce chien*) ;
- les syntagmes nominaux définis (p. ex. *le chien*, ou, en assimilant les déterminants possessifs aux déterminants définis, *son chien*) [6].

À ces catégories principales, s'ajoutent certaines unités lexicales isolées à l'intérieur d'une catégorie, comme, par exemple, certains adverbes ou l'adjectif *tel* :

(10) Ton père était mort SUR L'ÉCHAFAUD. [...] Je l'ai su après que ton père était mort *ainsi*...

(11) Juridiquement, la propriété civile est LA PART IDÉALE DE CHAQUE CO-PROPRIÉTAIRE DANS LA PROPRIÉTÉ COLLECTIVE DES CITOYENS. Quel sera le contenu réel d'*une telle part idéale*, c'est le marché qui le montrera.

Dans le cas de ces unités lexicales, le caractère anaphorique est étroitement lié au sens même du terme.

Des expressions pronominales et des ellipses, on peut dire qu'elles sont intrinsèquement sous-spécifiques. Les ellipses sont par définition perçues comme une absence d'expression et les expressions pronominales n'ont aucun contenu descriptif, en dehors d'une éventuelle information sur le genre et le nombre. L'interprétation des ces expressions *doit* donc, dans la très grande majorité des cas, être spécifiée par recours au contexte.

En ce qui concerne les syntagmes nominaux démonstratifs et définis, ce n'est pas seulement l'éventuelle sous-spécificité des expressions qui en fait des expressions potentiellement anaphoriques, mais aussi la nature du déterminant.

Selon Grevisse [37, §596], en employant un démonstratif, le locuteur « indique la situation dans l'espace (avec un geste éventuellement) de l'être ou de la chose désignés, ou parfois en les situant dans le temps ou dans le contexte ». C'est une partie inhérente du sens du démonstratif qu'il renvoie au contexte, celui-ci pouvant être le discours ou la situation d'énonciation. Les syntagmes nominaux démonstratifs seront donc presque toujours anaphoriques (renvoi au contexte textuel) ou déictiques (renvoi à la situation d'énonciation).

En employant un déterminant défini, le locuteur indique qu'il suppose connu de son interlocuteur l'être ou la chose qu'il désigne (voir [37, §564]). Cette présupposition que l'être désigné est connu intervient dans le fait que les syntagmes

[5] Les ellipses ne sont pas à proprement parler des expressions, mais plutôt des absences d'expression. On peut cependant les voir comme des occurrences de l'« expression vide ». Voir, par exemple, cette énumération de Ranta [75, p. 77] : « [Anaphoric expressions] include, in addition to pronouns, definite noun phrases, like *the man*, and modified definite phrases, like *the rich man*, but also the zero sign, that is, ellipsis. »

[6] Les analyses peuvent diverger sur l'inclusion ou non des syntagmes nominaux démonstratifs et définis parmi les expressions anaphoriques. Voir [23, p.37].

nominaux définis soient considérés comme anaphoriques. Le contraste entre les deux phrases suivantes illustrera ce point :

(12) a. Nous entrâmes dans un village ; l'église était sur une hauteur.

b. Nous entrâmes dans un village ; une église était sur une hauteur.

Les deux syntagmes *l'église* et *une église* peuvent être considérés comme également spécifiques. Cependant, dans la plupart des analyses, seul le syntagme *l'église* sera considéré comme anaphorique. Dans leur inventaire des facteurs constitutifs de l'anaphore associative, Kleiber et al. notent [54, p. 10] : « Le premier facteur est constitué par l'introduction d'un nouveau référent par l'expression en anaphore associative. [...] Précision capitale : cette introduction se fait sur le mode du connu (ou du « défini » ou encore du « déterminé »). ». Ainsi, en (12a), le syntagme *l'église* est une anaphore associative parce qu'il présuppose un référent unique et connu (l'église en question est l'église du village mentionné et est la seule église de ce village), alors qu'en (12b), le syntagme *une église* n'est pas une anaphore asssociative dans la mesure où le déterminant indéfini n'induit pas une telle présupposition. Ce qui fait donc d'un syntagme nominal défini une expression potentiellement anaphorique n'est pas seulement le fait qu'il soit sous-spécifique, mais qu'il soit sous-spécifique et désigne un objet supposé connu.

Au terme de cette présentation de la notion d'anaphore, récapitulons les deux points principaux :

- on rencontre dans les textes des expressions qui sont telles que leur interprétation nécessite qu'elles soient mises en relation avec une ou plusieurs autres expressions du même texte,
- et, dans le cas général, ces expressions sont caractérisables comme appartenant à certaines catégories.

1.2 Reprises

Notre objet d'étude central dans cette première partie de la thèse est ce que nous appelons les « phénomènes de reprise », phénomènes que nous caractérisons comme suit :

> Il y a reprise lorsqu'entre des expressions d'un même texte, non liées par les liens habituels de la syntaxe, il existe un lien sémantique qui doit être caractérisé en ayant recours à une relation d'identité.

Précision terminologique : le terme « reprise » est employé traditionnellement dans un sens très proche d'« anaphore » (voir [23]), si bien que les deux termes font en quelque sorte presque double emploi. Dans la mesure où nous avons besoin d'un nom pour les phénomènes que nous voulons délimiter (et qui à notre connaissance n'ont pas été délimités ainsi auparavant), nous nous permettons de donner ici à « reprise » une nouvelle acception, le terme « anaphore » étant

disponible pour décrire ce que décrit traditionnellement le terme « reprise ». Il faut noter que, dans notre optique, il n'y a pas quelque chose qui existerait indépendamment de l'observation et qui serait La Reprise, phénomène existant *a priori* ; il y a des phénomènes que l'on peut choisir de délimiter de telle ou telle manière et de nommer de telle ou telle manière. La multiplicité des points de vue sur un même objet ne devrait pas nuire à l'amélioration des observations sur cet objet.

Notre notion de reprise a en commun avec la notion d'anaphore le fait de décrire une relation entre des expressions qui ne sont pas liées syntaxiquement. En revanche, les notions de reprise et d'anaphore se distinguent l'une de l'autre sur deux points :

- la notion de reprise, contrairement à celle d'anaphore, ne fait pas intervenir une caractérisation du phénomène par la forme des expressions ;
- la notion d'anaphore, contrairement à celle de reprise, peut mettre en jeu des liens caractérisés sans recours à une relation d'identité.

La notion d'anaphore et celle de reprise couvrent deux ensembles de phénomènes qui ont une intersection non vide. Cette dernière est déterminée par deux lignes de partage :

- celle qui, à l'intérieur des phénomènes d'anaphore, distingue les anaphores caractérisées par recours à l'identité de celles qui ne le sont pas,
- et celle qui, à l'intérieur des phénomènes de reprise, distingue les reprises anaphoriques des reprises non anaphoriques.

Ces deux lignes de partage sont commentées dans les deux sous-sections suivantes. Les motivations qui nous ont conduit à les caractériser seront développées dans la section 1.3.

1.2.1 Liens caractérisés par ou sans recours à l'identité

Au-delà de la relation anaphorique, les différents exemples que nous avons présentés plus haut illustrent le fait que le lien qui permet l'interprétation de l'expression anaphorique peut prendre des formes différentes. Dans le même temps que nous caractérisons les différents types de liens en jeu dans ces exemples, nous déterminons deux ensembles de liens, selon qu'ils sont caractérisés *par recours* ou *sans recours* à une relation d'identité. Les premiers relèveront des phénomènes de reprise, les seconds non.

La notion de « recours à une relation d'identité » sera précisée par la suite à travers les différents exemples que nous examinerons.

Liens avec identité

En (1), les expressions *Jacques* et *Il* sont interprétées comme désignant une seule et même personne. Le lien entre les deux expressions est ici une identité de

l'objet désigné, c'est-à-dire ce que nous appellerons une « identité de dénotation ».

D'une manière analogue, en (2), l'expression *Les deux établissements* est interprétée comme désignant un ensemble de deux sociétés dont l'une est désignée par l'expression *Fortis* et l'autre par l'expression *la Générale de Banque*. L'ensemble constitué de la société désignée par *Fortis* et de la société désignée par *la Générale de Banque* est le même que celui qui est désigné par *Les deux établissements*.

Dans l'exemple (3), en revanche, l'expression anaphorique *celles de la Méditerranée* et sa source, *les plages de l'Atlantique*, désignent deux choses différentes, mais qui sont de même type (les deux choses désignées sont des plages). La relation est la même entre *les fleurs blanches* et *les rouges* dans l'exemple (4). Nous sommes là dans un cas que nous caractériserons comme une « identité de description ». Dans chacun de ces deux exemples, une même description (*plages* en (3), *fleurs* en (4)) s'applique à l'objet désigné par l'expression source et l'expression anaphorique.

Dans l'exemple (5), l'expression anaphorique *trois* et sa source *ces quatre administrateurs* désignent également deux « choses » différentes, mais qui sont de même type. Il y a ici, comme en (3) et (4), identité de description : *ces quatre administrateurs* désigne un ensemble d'administrateurs et *trois* également. De plus, l'ensemble désigné par *trois* est un sous-ensemble du premier ensemble d'administrateurs évoqué.

Dans le sens où ils mettent en jeu une relation d'identité, les liens anaphoriques que nous avons décrits ici relèvent de notre notion de reprise.

Liens sans identité

Si dans les exemples précédemment examinés, le lien entre l'expression anaphorique et sa source met en jeu une relation d'identité de la chose désignée ou du type de la chose désignée, ce n'est pas le cas dans les trois autres exemples donnés au début de la section 1.1. Ces exemples sont des cas de ce qu'il est d'usage d'appeler « anaphore associative ».

En (6), l'expression *le câble d'alimentation* ne désigne pas n'importe quel câble, mais le câble d'alimentation de l'imprimante dont on vient de parler, c'est-à-dire une partie de l'imprimante. En (7), l'objet désigné par *l'église* est caractérisé comme étant l'église du village mentionné par ailleurs. En (8), l'ensemble de personnes désigné par *les victimes* est caractérisé par rapport au tremblement de terre évoqué.

Dans ces trois exemples, le lien entre l'expression anaphorique et sa source n'est selon nous pas caractérisé en ayant recours à une relation d'identité. Dans les trois cas, l'expression anaphorique et sa source désignent deux êtres différents entre lesquels il n'y a pas identité de description : une imprimante n'est pas un câble, et inversement, un câble n'est pas une imprimante ; de même pour l'église et le village et pour le tremblement de terre et les victimes. On peut dire qu'il y a ici une relation qui permet de spécifier l'identité de l'objet désigné par l'expression

anaphorique, mais ce n'est pas là avoir recours à une *relation* d'identité, celle-ci étant entendue dans un sens comparatif, qui se traduit par l'usage de l'adjectif *même* dans nos description des exemples (1) à (5).

Dans ce sens, les liens anaphoriques que nous avons décrits ici ne relèvent pas de notre notion de reprise.

1.2.2 Liens de reprise non anaphoriques

Dans l'ensemble des phénomènes de reprise, on peut vouloir distinguer ceux qui relèvent de l'anaphore (p. ex. les exemples (1) à (5) ci-dessus) de ceux qui n'en relèvent pas. Notre objectif n'est cependant pas de caractériser cette ligne de partage. Au contraire, nous voulons mettre en avant l'idée qu'on peut en faire abstraction : les liens anaphoriques qui sont caractérisés par recours à l'identité peuvent également être observées dans des contextes qu'il n'est pas d'usage de qualifier d'« anaphore ».

L'exemple le plus évident est la relation de « coréférence », dont le texte suivant donne un exemple :

(13) Les critiques pleuvent. *M. Guilhaume*$_1$ a trahi l'esprit de la loi. *Il* a concentré dans *ses* mains tous les pouvoirs, alors qu'on ne *lui* demandait que d'harmoniser stratégies et programmes. *Il* a nommé dans chaque chaîne une armée mexicaine de responsables qui paralyse toute décision. *Il* traite à la légère les producteurs indépendants et *se* comporte comme un actionnaire irresponsable vis-à-vis de la Société française de production. *M. Guilhaume*$_2$ supprime des émissions sans même prévenir leurs responsables, improvise d'autres programmes dont l'audience est catastrophique. *Il* met en péril les ressources publicitaires, creuse les déficits d'A2 et de FR3, provoque la démission de leurs directeurs financiers. *M. Guilhaume*$_3$ cacherait, sous de beaux discours, une absence totale de projet pour la télévision publique.

Les expressions en italiques dans ce texte désignent toutes la même personne : M. Guilhaume ; en ce sens, on dit d'elles qu'elles sont « coréférentes ». Traditionnellement, certaines de ces expressions seront considérées comme anaphoriques (les formes pronominales), d'autres non (les différentes occurrences de l'expression *M. Guilhaume*). La notion de reprise dépasse cette distinction. Nous dirons que, hormis la première occurrence de l'expression *M. Guilhaume*, toutes les expressions en italiques dans ce texte sont des reprises.

Les exemples (3) et (4) mettaient en jeu des expressions désignant des objets différents mais de même type. D'une manière assez triviale, on rencontrera ce même type de lien entre deux expressions sans qu'il y ait anaphore. Dans l'exemple suivant :

(14) Marie aime les plages de l'Atlantique ; Pierre préfère les plages de la Méditerranée.

on considère qu'il y a un lien de reprise entre *les plages de la Méditerranée* et *les plages de l'Atlantique*.

Dans l'exemple (5), l'expression anaphorique *trois* désigne un sous-ensemble de l'ensemble désigné par *les quatre administrateurs*. Là encore, on pourra trouver ce même type de relation entre deux expressions sans qu'il y ait à proprement parler anaphore. Dans la phrase suivante :

(15) Quatre-vingts chevreuils ont été lâchés en un même point d'un plateau calcaire des Monts du Vaucluse, au sud de Carpentras, couvert de chênes verts (Quercus ilex), de chênes pubescents (Quercus pubescens) et de conifères (cèdre bleu replanté). Dix animaux étaient équipés de colliers émetteurs.

le syntagme nominal indéfini *Dix animaux* (*a priori* non anaphorique) désigne un sous-ensemble de l'ensemble désigné par *Quatre-vingts chevreuils*. La relation d'identité est ici à la fois une identité de description (les dix animaux en question sont des *chevreuils*), et une identité partielle au niveau des objets désignés (les dix animaux sont dix des quatre-vingts chevreuils).

À travers ces différents exemples, on voit que notre notion de reprise n'a en commun avec la notion d'anaphore que l'idée d'une relation entre expressions, sans poser de condition sur la forme des expressions qui sont liées. Dans ce sens, la notion de reprise est une généralisation de la notion d'anaphore, dans les cas où cette dernière met en jeu une relation d'identité.

1.2.3 Précisions

Avant d'aborder les motivations qui nous ont conduit à adopter la notion de reprise que nous avons définie, nous apportons ici deux précisions.

Comme un lien anaphorique, un lien de reprise est une relation entre une expression et une ou plusieurs autres expressions, dont nous dirons qu'elles sont les « sources » de la « reprise ». La relation de reprise est donc pour nous orientée :

- lorsqu'un lien de reprise est un lien anaphorique, l'orientation de la relation est celle du lien anaphorique : elle va de l'expression la moins spécifique (c'est-à-dire l'expression anaphorique) vers l'expression la plus spécifique. ;
- lorsqu'un lien de reprise n'est pas anaphorique, la relation est orientée suivant la linéarité du texte : si deux expressions e_i et e_j sont liées par un lien de reprise et si e_i précède e_j, alors e_i est la source de e_j.

Par ailleurs, rappelons que la relation de reprise, comme la relation anaphorique, vise à décrire des liens entre expressions qui ne sont pas explicités dans le discours par les liens habituels de la syntaxe.

1.3 Motivations

L'intérêt de la notion de reprise par rapport à celle d'anaphore est pour nous double. D'une part, elle permet de se concentrer sur le résultat de l'interprétation, plutôt que sur le processus aboutissant à l'interprétation. D'autre part, elle permet de restreindre le champ d'observation à un ensemble de données sur lesquelles l'inter-subjectivité des observations devrait pouvoir, selon nous, être établie. Nous développons ces deux idées dans les deux sections suivantes.

1.3.1 Une approche centrée sur le résultat de l'interprétation

Pour caractériser notre approche, nous la mettons ici en perspective avec une distinction faite par F. Corblin, dans *Les formes de reprise dans le discours* [23, p. 177]. Corblin oppose anaphore et coréférence, sur la base du calcul mis en jeu dans l'interprétation des expressions : « L'opposition la plus nette est certainement celle qui vaut entre l'identité de référence (ou co-référence), et l'anaphore [7]. » À titre d'exemple, Corblin commente le texte suivant [8] :

> Une des plus belles œuvres de **Segalen** est certainement comme c'est le cas pour *Flaubert* **sa** correspondance (en très grande partie inédite). Souvent éloigné de **ses** amis les plus proches, voire de **sa** femme, **Segalen** leur écrivait et, tout comme *l'ermite de Croisset* livrait les secrets de *son* œuvre, levait le voile qui masquait certaines régions de **lui-même**.
>
> (Claude Courtot, *Victor Segalen*, Henri Veyrier, p. 30)

Dans ce texte, Corblin considère que, d'une part, la relation de *sa*, dans *sa correspondance*, *ses*, dans *ses amis*, et *sa*, dans *sa femme*, vers la première occurrence de *Segalen* relève de l'anaphore et que, d'autre part, la relation de la seconde occurrence de *Segalen* vers la première relève de la coréférence, de même que la relation de *l'ermite de Croisset* vers *Flaubert*.

La distinction entre les deux concepts, également formulée comme une distinction entre « identité d'interprétation et interprétation par reprise » [23, p. 111], est justifiée par une différence dans le calcul qui amène à identifier « les relations d'identité dont les expressions linguistiques sont le support » [9] :

> Il y a identité d'interprétation si a et b reçoivent la même interprétation en vertu de règles qui ne doivent rien à leur proximité dans le même segment linguistique ; cela s'applique à deux occurrences d'un même nom propre ou de la même unité lexicale, au couple formé d'un

[7] Chapitre 7. « Les chaînes de référence naturelles ». p. 177. Voir aussi, entre autres, le chapitre « Introduction », p. 31–32, et le chapitre 4, « L'anaphore nominale », p. 111–112.

[8] En **gras** les références à Segalen, en *italiques* les références à Flaubert.

[9] Corblin signale que « la différence entre ces deux relations a été notée par M. Gross (1973) et formulée dans ces termes par J.-C. Milner (1982) ».

nom propre et d'une description définie identifiante (*Aristote / Le maître d'Alexandre*), à deux occurrences de *je*, etc.

Il y a interprétation par reprise [10] si un terme, b, exige pour être interprété l'emprunt à un terme proche a d'un élément qui fixe l'interprétation de b : cela s'applique par exemple aux couples dont le second terme est un pronom.

La distinction effectuée par Corblin vise à caractériser l'emploi de différentes expressions. Nous ne mettons pas en doute qu'elle soit motivée, mais elle n'est pas pertinente au niveau de description des données où nous nous plaçons. Dans la perspective de l'implantation d'un système qui caractérise les différents êtres désignés par un texte et les différentes mentions de ces êtres dans ledit texte, toutes les expressions en gras dans le texte ci-dessus devront être associées à la personne appelée « Segalen » et toutes les expressions en italiques à la personne appelée « Flaubert », qu'elles soient anaphoriques ou non. Nous nous contenterons donc de décrire la relation entre ces expressions comme une relation de coréférence [11].

Pour caractériser notre approche, nous dirions que, dans une certaine mesure, nous nous focalisons sur le résultat que devra obtenir notre éventuel interpréteur de texte et non sur le processus qui permettra d'obtenir ce résultat. Dans le cas des expressions du texte cité par Corblin, la relation de coréférence est celle que devra identifier le système que nous implanterons éventuellement ; elle est donc celle que, pour l'heure, nous voulons observer. Si des distinctions au niveau du processus en œuvre dans l'interprétation de tel ou tel type d'expression existent, leur modélisation sera éventuellement du ressort des règles implantées dans notre interpréteur de textes.

En résumé, notre système informatique devra spécifier les différents éléments des univers dont parlent les textes qu'on lui donnera à analyser et l'accomplissement de cette tâche passe par l'identification d'un certain nombre de liens entre les expressions des textes en question, parmi lesquels figurent les liens de reprise qui font l'objet de notre attention. Nous devons pour l'heure spécifier la sortie du système, ce qui sera fait à travers une typologie des liens de reprise présentée au chapitre 2.

1.3.2 Des observations faisant l'objet d'une inter-subjectivité

On pourra juger la notion de reprise relativement restrictive dans le sens où elle se limite à la caractérisation de liens mettant en jeu une relation d'identité. Par ailleurs, on pourra s'interroger sur les motivations qui nous conduisent à exclure les phénomènes d'anaphore associative. La généralisation que représente

[10] Attention : nous ne donnons pas au terme « reprise » la même extension que celle que lui donne Corblin. Chez ce dernier, « reprise » est globalement équivalent à « anaphore ».

[11] Cet emploi du terme « coréférence » est d'ailleurs tout à fait courant en linguistique informatique. Voir, par exemple, les campagnes d'évaluation MUC [44].

dans une certaine mesure la notion de reprise par rapport à celle d'anaphore (c'est la relation qui compte, non la forme de l'expression) ne serait-elle pas applicable aussi aux phénomènes d'anaphore associative ?

La raison pour laquelle nous restreignons la notion de reprise à des liens entre expressions caractérisés en ayant recours à une relation d'identité est principalement que nous pensons que ces liens sont les plus susceptibles d'être observés de manière inter-subjective, c'est-à-dire que différents observateurs observeront bien les mêmes liens de reprise dans les mêmes textes.

La problématique de l'inter-subjectivité des observations sera abordée au chapitre 4. Nous y présentons une expérience dans laquelle différents observateurs devaient noter les liens qu'ils observaient dans les textes, suivant une version plus ancienne de la typologie que nous présentons aujourd'hui. Cette version incluait dans les phénomènes à observer des cas relevant de l'anaphore associative, élargie à une absence de restriction sur le type de syntagme nominal en jeu. Les résultats obtenus montrent clairement une absence d'inter-subjectivité sur les phénomènes qui ne relèvent pas de la coréférence au sens strict. Ils nous ont donc conduit à réviser ou raffiner les définitions que nous avions mises en place à l'époque, pour aboutir à la notion de reprise que nous proposons aujourd'hui. Nous pensons pouvoir attester, dans l'avenir, que cette nouvelle notion de reprise donne lieu à des observations inter-subjectives.

1.4 Terminologie

À la base de la caractérisation des différents types de liens de reprise que nous décrirons par la suite se trouvent les concepts de « dénotation » et « description ». De manière générale, la distinction est celle qu'il est d'usage de faire entre les mots (« descriptions ») et les choses désignées par les mots (« dénotation »).

Lorsqu'on touche à la question des objets désignés par les expressions, on emploie en général des termes tels que « dénoter », « désigner », « faire référence » ou leurs dérivés (p. ex. « expressions référentielles », « coréférence »). L'acception donnée à ces différents termes peut cependant varier d'un auteur à l'autre, ou d'un domaine à l'autre. En particulier, pour certains, on dira d'une expression qu'elle « fait référence » à un objet ou « dénote » un objet seulement si cet objet existe dans la réalité (voir la citation de Reboul ci-dessous), alors que pour d'autres, la notion de référence sera plus attachée à l'emploi de certaines expressions, qui peuvent éventuellement désigner des êtres imaginaires. Par exemple, pour Ducrot [31, p. 317] :

> La communication linguistique ayant souvent pour objet la réalité extra-linguistique, les locuteurs doivent pouvoir désigner les objets qui la constituent : c'est la **fonction référentielle** du langage (le ou les objets désignés par une expression forment son référent). Cette réalité

> n'est cependant pas nécessairement *la* réalité, *le* monde. Les langues naturelles ont en effet ce pouvoir de construire l'univers auquel elles se réfèrent ; elles peuvent donc se donner un **univers de discours** imaginaire. L'île au trésor est un objet de référence possible aussi bien que la gare de Lyon.

La terminologie utilisée par Reboul et al. [76] illustre bien ces deux aspects (renvoi à la réalité et usage de certaines expressions) donnés à la notion de « référence ». On en arrive en effet à des objets étranges tels que des « expressions référentielles non référentielles » :

> Il peut paraître étrange d'intituler un paragraphe *ER (expressions référentielles) référentielles et non référentielles*, mais il faut cependant admettre qu'il y a des ER qui ne réfèrent pas. On peut penser ici aux noms qui renvoient à des personnages de fiction ou à des êtres mythologiques (*le Père Noël, Sherlock Holmes, Pégase*, etc.). Il faut aussi penser aux descriptions indéfinies (ex : *un chat*), dont on s'accorde généralement à penser qu'elles ne réfèrent pas.

Notre objectif n'est pas ici de faire l'étude des différentes visions des notions de dénotation et référence et de discuter les motivations qui conduisent tel ou tel auteur à adopter tel point de vue. Nous nous contenterons de fixer dans les sous-sections qui vont suivre la terminologie que nous utiliserons dans la thèse. Cette terminologie est totalement déterminée par nos objectifs, en l'occurrence la typologie des reprises présentée au chapitre 2 et, dans une perspective plus lointaine, l'implantation d'un système d'interprétation automatique des textes. Dans ce contexte, le point important est d'identifier ce dont parlent les textes, indépendamment de la question de savoir s'il s'agit de la réalité ou non. Nous adopterons donc un point de vue qui consiste à essayer de déterminer quels sont les objets dont parle un texte *à partir des expressions de ce texte* et non en fonction de leur existence ou non-existence en dehors du texte.

1.4.1 Univers de dénotation

Nous supposerons qu'à tout texte est associé ce que nous appellerons un « UNIVERS DE DÉNOTATION », univers qui contient les objets, personnes, événements, faits, situations, etc. dont parle le texte. L'univers de dénotation, dans notre terminologie, est ce que Dowty et al. [30] appellent « le vaste complexe de choses dont les textes peuvent parler. »

Nous dirons des objets, personnes, événements, faits, situations, etc. dont parle le texte, qu'ils sont des « ÊTRES » de l'univers de dénotation. Les êtres de l'univers de dénotation sont *censés être réels* à partir du moment où les expressions du texte les désignent. La caractérisation de l'univers de dénotation associé à un texte dépend donc des expressions de ce texte.

Nous utiliserons également le terme « RÉFÉRENTS » pour parler des êtres de l'univers de dénotation.

1.4.2 Expressions dénotantes

On appelle « expression dénotante » une expression qui dénote un être de l'univers de dénotation. À l'intérieur de la classe des expressions dénotantes, nous distinguons deux sous-classes :

- les expressions référentielles, qui désignent un être supposé existant en dehors du discours ou identifiable dans le discours,
- et les expressions référentes, qui introduisent un nouvel être dans le discours, être qui n'est pas supposé connu.

Nous caractérisons ci-dessous ces deux types d'expressions dénotantes. Signalons qu'aux expressions dénotantes s'opposeront dans la section suivante (1.4.3) des expressions *non dénotantes*.

EXPRESSIONS RÉFÉRENTIELLES. Nous caractérisons les expressions qui désignent un être supposé existant en dehors du discours en reprenant la taxonomie proposée par Ducrot [31, p. 320]. Ces expressions seront donc pour nous :

- les « descriptions définies », c'est-à-dire les syntagmes nominaux déterminés par un article défini ou un déterminant possessif (p. ex. *le chien, son chien*), auxquels nous adjoignons les pronoms « définis » et les déterminants possessifs de troisième personne [12] ;
- les noms propres (p. ex. *Dieu, Rabelais*) ;
- les démonstratifs, c'est-à-dire, de manière générale, les pronoms *ce, ceci, cela, celui, celle,* etc. et les syntagmes nominaux déterminés par *ce, cet, cette, ces* [13] ;
- les déictiques, « expressions dont le référent ne peut être déterminé que par rapport aux interlocuteurs ». Cette classe inclut les pronoms de première et deuxième personnes et des adverbes tels que *ici, hier, aujourd'hui, demain*.

On notera qu'une expression dénotante n'appartenant pas à l'une de ces catégories peut être une expression référentielle si elle désigne un être déjà introduit dans le discours par une autre expression dénotante.

EXPRESSIONS RÉFÉRENTES. Pour caractériser les expressions référentes (c'est-à-dire qui introduisent un nouvel être dans l'univers de dénotation), nous utilisons un critère proposé à l'origine par Karttunen [51]. L'objectif de Karttunen dans l'article en question était de caractériser les différents contextes dans lesquels un

[12] Nous appelons « pronoms définis » les pronoms personnels (p. ex. *il, lui, elles*) et les pronoms relatifs (p. ex. *qui, que, dont*), ces termes devant être entendus selon la classification de Grevisse [37, 2e partie, chapitre IV].

[13] Ducrot ajoute à ces cas certains emplois de l'article défini.

syntagme nominal indéfini introduit ou pas un « référent de discours [14] ». Il le fait en se basant sur le fait que « les descriptions définies expriment une présupposition existentielle : appeler quelque chose « *le* ... », c'est présupposer que cette chose doit exister [15]. » À partir de là, Karttunen expose le critère suivant : « Accordons-nous pour dire que l'apparition d'un syntagme nominal indéfini établit un « référent de discours » dans le cas où elle justifie l'occurrence d'un pronom coréférentiel ou un syntagme nominal défini dans la suite du discours. »

De manière similaire, nous dirons d'une expression e_i, qui n'est pas une expression dénotante référentielle, qu'elle est une expression dénotante référente s'il est possible de poursuivre le discours où elle apparaît en utilisant un syntagme nominal défini ou un pronom défini [16] qui fasse référence à l'être désigné par l'expression e_i. Ainsi, dans l'exemple (16), reproduit ici :

(16) Pierre cherche une licorne, mais il voudrait qu'elle ne soit pas blanche.

nous dirons que le syntagme *une licorne* est une expression référente qui désigne un être de l'univers de dénotation, dans la mesure où nous interprétons le pronom *elle* comme désignant la licorne que cherche Pierre.

On notera que le critère utilisé par Karttunen pour les syntagmes nominaux indéfinis peut s'appliquer à d'autres expressions, en particulier une proposition. Par exemple, dans le texte suivant,

(17) Après avoir cédé sa banque d'affaires à la Société Générale, le groupe britannique Hambros a vendu sa participation de 52 % dans Hambro Insurance Services Group au canadien Lindsey Morden Groupe. Le montant de la transaction est de 44,9 millions de livres...

nous considérons que le syntagme *la transaction* fait référence à la vente évoquée dans la phrase précédente. Nous dirons donc que la proposition *le groupe britannique Hambros a vendu sa participation de 52 % dans Hambro Insurance Services Group au canadien Lindsey Morden Groupe* est une expression dénotante, qui introduit dans l'univers de dénotation l'être auquel *la transaction* fait référence, c'est-à-dire la vente en question.

[14] Le terme « référent de discours » n'est pas très précisément défini dans [51]. Nous l'interprétons, dans l'article en question, comme caractérisant ce que nous appelons un « être de l'univers de dénotation » *lorsque celui-ci est introduit par le discours*, par opposition aux objets du monde réel auquel le discours est susceptible de faire référence (ceux-ci étant des référents, tout court). Depuis [51], le terme « référent de discours » a été souvent employé pour désigner ce que nous appelons ici « êtres de l'univers de dénotation », c'est-à-dire tout objet dont un texte est susceptible de parler, qu'il soit réel ou non, supposé existant dans la situation dénonciation ou non. Par ailleurs, dans le contexte de la DRT, un « référent de discours » est un objet formel du langage utilisé pour représenter le discours. Face à ces ambiguïtés potentielles, le terme « être de l'univers de dénotation », que nous avons défini, a l'avantage, pensons-nous, de ne pas prêter à confusion.

[15] Dans le même ordre d'idée, Grevisse [37, §564] dit : « L'article défini s'emploie devant le nom qui désigne un être ou une chose connus du locuteur et de l'interlocuteur. »

[16] Voir la note 12.

DÉNOTER, DÉSIGNER, FAIRE RÉFÉRENCE. Nous avons employé plus haut les expressions *désigner un être de l'univers de dénotation* ou *faire référence à un être de l'univers de dénotation*. Nous emploierons les trois verbes *dénoter, désigner* et *faire référence* de la manière suivante. De manière générale, pour l'ensemble des expressions dénotantes, qu'elles soient référentes ou référentielles, nous dirons :

– une expression dénotante « désigne » ou, de manière équivalente, « dénote » un être de l'univers de dénotation.

De façon plus spécifique, pour les expressions dénotantes référentielles, nous dirons aussi :

– une expression référentielle « fait référence » à un être de l'univers de dénotation.

1.4.3 Descriptions et expressions non dénotantes

De manière générale, nous dirons qu'un discours non seulement désigne des êtres, mais aussi qu'il décrit ces êtres. Les expressions d'un texte fournissent donc des « descriptions » des êtres de l'univers de dénotation associé au texte.

En règle générale, toutes les expressions du langage, à l'exception des pronoms, contribuent à la description des êtres de l'univers de dénotation. Une expression dénotante a ainsi aussi en général une fonction descriptive. Dans la phrase :

(18) Pierre mange une pomme.

les expressions *Pierre* et *une pomme* désignent chacune un être de l'univers de dénotation, mais elles décrivent aussi dans le même temps ces êtres comme s'appelant *Pierre* et comme étant une *pomme*, respectivement.

Nous analysons cependant certaines expressions comme ayant dans les textes seulement une fonction descriptive, c'est-à-dire comme n'étant pas des expressions dénotantes.

Par exemple, dans la phrase :

(19) Pierre est un imbécile.

nous considérons que l'expression *Pierre* désigne un être de l'univers de dénotation, mais pas l'expression *un imbécile*. Nous disons de celle-ci qu'elle ne fait que décrire l'être auquel *Pierre* fait référence.

Dans cette phrase, le syntagme *un imbécile* a une fonction d'attribut. De manière générale, les expressions qu'on décrit traditionnellement comme ayant fonction d'attribut (voir [37, §241-251]) seront pour nous des expressions non dénotantes. Il en est de même pour les expressions épithètes. Dans l'expression *un crayon rouge*, nous ne considérons pas qu'il y a une désignation de la couleur rouge. L'adjectif *rouge* vise seulement à indiquer une propriété du crayon dénoté par le syntagme *le crayon rouge*.

Les syntagmes verbaux constituent aussi une catégorie d'expressions que nous considérons comme non dénotantes. Dans sa description des éléments fondamentaux de la phrase, Grevisse caractérise dans un premier temps les notions de sujet et prédicat et dit du sujet qu'il « représente ce dont je dis quelque chose » et du prédicat (*a priori* un syntagme verbal) qu'il « représente ce que je dis [du sujet] » [37, §226]. Dans l'exemple (18) ci-dessus, nous considérons que le syntagme verbal *mange une pomme* ne fait que dire quelque chose sur l'être dénoté par *Pierre* ; il n'introduit par dans l'univers de dénotation un être qui serait le fait de manger une pomme en général.

1.4.4 Êtres singuliers et ensembles d'êtres singuliers dans l'univers de dénotation

Nous terminons cette présentation des quelques notions préliminaires à la présentation de notre typologie des reprises en caractérisant deux types d'êtres de l'univers de dénotation : les « êtres singuliers » et les êtres qui sont des « ensembles d'êtres singuliers ».

Dans la phrase :

(20) Les cinq groupes intéressés par l'assureur public étaient au rendez-vous pour rendre leur copie au ministère des Finances.

nous dirons que *les cinq groupes intéressés par l'assureur public* dénote un être qui est un ensemble de groupes, et que *l'assureur public* dénote un être qui au contraire est un être singulier.

Les notions d'« être singulier » et d'« ensemble d'êtres singuliers » dans notre terminologie correspondent respectivement à un atome et à un ensemble *d'au moins deux éléments* en théorie des ensembles.

Un ensemble d'êtres singuliers est un être de l'univers de dénotation au même titre qu'un être singulier. Étant donné la phrase :

(21) Jacques a rejoint Pierre au café, puis ils sont partis au cinéma.

nous considérons que les expressions *Jacques* et *Pierre* dénotent chacune un être singulier et que le pronom *ils* dénote un troisième être, qui est un ensemble d'êtres singuliers constitué des êtres Jacques et Pierre.

Parmi les ensembles d'êtres singuliers, on distingue les classes d'êtres singuliers. Lorsqu'une expression dénote une classe d'êtres de l'univers de dénotation, nous parlons de « dénotation générique ». Soit les phrases :

(22) a. Le lion s'est échappé du zoo.

b. Le lion est un félin.

Le syntagme *Le lion* dans la première phrase désigne un être particulier de l'univers de dénotation : un lion particulier qui s'est échappé d'un zoo particulier. Dans la seconde phrase, par contre, ce même syntagme ne désigne pas un être

particulier ; ce que dit cette phrase est que tout être de l'univers de dénotation qui est un lion est un félin.

Nous dirons que l'exemple (22b) est un cas de dénotation générique. Le syntagme *Le lion* dans cette phrase fait référence à une classe (ou à un genre) d'êtres de l'univers de dénotation en général, plutôt qu'à un être particulier de l'univers de dénotation. Comme tout ensemble d'être singuliers, une classe d'êtres de l'univers de dénotation est elle-même un être de l'univers de dénotation.

Nous avons vu que les êtres de l'univers de dénotation sont pour nous censés exister à partir du moment où une expression les désigne. Dans le cas des classes, « exister dans l'univers de dénotation » signifie n'être pas vide dans cet univers, ce qui, l'univers de dénotation n'étant pas nécessairement le monde réel, ne préjuge en rien de l'existence ou non dans la réalité d'êtres qui appartiennent à la classe.

Ainsi, étant donné la phrase :

(23) Les licornes n'existent pas.

nous dirons que l'expression *Les licornes* fait référence à une classe d'êtres de l'univers de dénotation associé à cette phrase.

La phrase en (23) peut être analysée comme contenant un sous-entendu :

(23) Les licornes n'existent pas *dans la réalité*.

On peut voir la réalité comme l'ensemble des choses, événements, etc. réels. Dans ce sens, la phrase en (23) fait référence à deux ensembles, l'ensemble R des choses réelles et l'ensemble L des licornes. La manière dont nous voyons l'univers de dénotation associé à cette phrase est représentée figure 1.1. L'univers de dénotation inclut à la fois l'ensemble des choses réelles et l'ensemble des licornes et la phrase dit que ces deux ensembles sont disjoints.

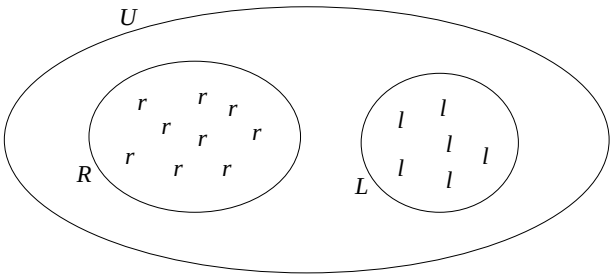

FIG. 1.1 – Univers de dénotation associé à la phrase *Les licornes n'existent pas*.

Les éléments de R et L sont représentés par r et l, respectivement. On notera que l'ensemble L n'est pas vide ; parler des licornes, c'est supposer leur existence, au moins en tant qu'êtres imaginaires. L'ensemble L est en dehors de la réalité, mais à l'intérieur de l'univers de dénotation U.

1.5 Notation des observations

Après avoir posé les grandes lignes de notre champ d'observation et fixé le sens des termes que nous utiliserons, il nous reste à nous donner un système de notation des observations que nous ferons. L'intérêt d'un tel système est double : d'une part, il permettra de noter les observations de manière concise, d'autre part, nous pourrons demander à divers observateurs de noter leurs observations avec ce système de notation, de manière à pouvoir évaluer si des observateurs différents font bien les mêmes observations. Le but sera alors de démontrer que la typologie des reprises que nous avons définie est opérationnelle. Si tel est le cas, on aura ainsi attesté qu'un système d'identification automatique des liens de reprise peut effectivement être évalué.

L'approche développée dans cette première partie de la thèse est donc globalement la même que celle qui est adoptée dans les différents projets d'annotation de corpus (en particulier les conférences MUC [38], pour lesquelles une évaluation de l'inter-subjectivité des observations a effectivement été menée [20]).

Notre système de notation des observations est présenté ici dans ses grandes lignes. Plus de détails seront donnés par la suite avec la présentation de chacun des différents types de reprises.

1.5.1 Délimitation des expressions

Les observations que nous aurons à faire mettront en jeu des relations entre expressions. La première étape consistera à délimiter ces expressions par des chevrons [17]. Par exemple, dans l'exemple (1), reproduit ici, le lien de reprise met en jeu les deux expressions *Jacques* et *Il*. On place ces deux expressions entre chevrons :

(1) <Jacques> dort. <Il> est fatigué.

La plupart du temps, les expressions à annoter seront des syntagmes nominaux. Pour alléger la notation en évitant au maximum les enchâssements, on n'annotera que les syntagmes nominaux noyau [18], étant entendu que l'interprétation du syntagme se fera toujours en tenant compte des compléments éventuels à droite du noyau. Par exemple, dans l'exemple (6), reproduit ci-dessous, les expressions pour lesquelles on observe un lien de reprise sont *l'imprimante* et *le*

[17] On aurait pu adopter un système de notation dans les langages d'annotation classiques que sont SGML et XML, mais il nous a semblé que le système d'annotation que nous utilisons faciliterait la lisibilité des exemples. De fait, le langage de marquage proposé ici ne vise pas à être un format d'échange en dehors de la présente thèse. En marge de la thèse, nous avons participé à un projet d'annotation des anaphores « grammaticales » dans un corpus de taille importante [90]. Le schéma d'annotation proposé dans ce projet donne une idée du schéma XML que nous adopterions.

[18] Un syntagme nominal noyau est la partie d'un syntagme nominal qui va de son début à son noyau (ou « tête ») inclus.

câble d'alimentation ; on n'annotera que le syntagme noyau pour cette dernière expression :

(6) Ne placez pas <l'imprimante> à un endroit où des personnes pourraient marcher sur <le câble> d'alimentation.

Même principe pour l'exemple (3), où les deux expressions en jeu sont *les plages de l'Atlantique* et *celles de la Méditerranée* : on annotera seulement le syntagme noyau *les plages* et le syntagme noyau *celles*.

(3) Marie aime <les plages> de l'Atlantique ; Pierre aime <celles> de la Méditerranée.

1.5.2 Description du lien observé

Après avoir délimité les expressions observées comme étant en relation, on insérera entre crochets angulaires (`[` et `]`) immédiatement avant le chevron fermant pour chacune des expressions les informations nécessaires à la description de la relation en jeu. Soit, par exemple :

(1) <Jacques `[xxx]`> dort. <Il `[xxx]`> est fatigué.

où `xxx` sera remplacé par les informations nécessaires à la description du lien observé. Ces informations seront présentées progressivement avec la description détaillée des différents types de liens de reprise, description qui fait l'objet du chapitre suivant.

Chapitre 2

Une typologie des liens de reprises

Nous avons défini au chapitre précédent la notion de reprise dans les termes suivants :

> Il y a reprise lorsqu'entre des expressions d'un même texte, non liées par les liens habituels de la syntaxe, il existe un lien sémantique qui doit être caractérisé en ayant recours à une relation d'identité.

Le présent chapitre propose une typologie des différents liens de reprise qu'on peut observer dans les textes. On distingue cinq types de lien de reprise :

- l'identité de dénotation, ou coréférence : les expressions reliées désignent le même être de l'univers de dénotation ;
- l'identité de description : deux expressions dénotantes différentes désignent des êtres différents mais de même type, ou deux expressions non dénotantes sont interprétées comme véhiculant la même description ;
- la relation `membre-de`, qui décrit les cas où une expression dénotante désigne un élément ou un sous-ensemble d'un ensemble désigné par une autre expression dénotante ;
- la relation `distingué-de`, qui met en jeu l'usage de l'adjectif *autre*, qui exprime la négation de l'identité entre deux êtres, tout en les envisageant comme des êtres de même type ;
- et enfin, un type de reprise plus particulier qui met en jeu un pronom objet faisant référence à un discours.

Ces cinq types de reprise sont décrits en détail dans les sections 2.1 à 2.5.

En complément à cette typologie des reprises, nous évoquons dans la section 2.6 les phénomènes d'anaphore ou de deixis qui ne sont pas couverts par la notion de reprise. Enfin, dans une dernière section, nous mettons en perspective

notre travail par rapport à quelques approches des problèmes abordés. Nous distinguerons notre approche de celle qui est adoptée en sémantique formelle, puis nous comparerons notre typologie des reprises aux descriptions proposées par M. Halliday et R. Hasan [43], au schéma d'annotation de la « coréférence » proposé dans le projet MATE [69] et aux descriptions proposées par S. Salmon-Alt [82, 81].

2.1 Identité de dénotation

La relation de reprise qu'on observera le plus souvent dans les textes est l'identité de dénotation ou coréférence.

2.1.1 Définition

On dit qu'il y a identité de dénotation entre deux expressions lorsqu'elles dénotent le même être de l'univers de dénotation. L'identité de dénotation est donc une relation entre des expressions dénotantes.

La notion d'identité de dénotation que nous utilisons ici correspond globalement à celle de « coréférence » telle qu'elle est utilisée dans le *Coreference Task* de MUC [44] [1]. Nous utiliserons donc par la suite les termes « identité de dénotation » et « coréférence » de manière interchangeable.

De deux expressions liées par une identité de dénotation, nous dirons qu'elles sont « coréférentes».

La notion de coréférence n'est pas nouvelle et nous en avons déjà vu des exemples au chapitre 1 (voir en particulier l'exemple (13) page 28). La notion de coréférence a cependant pu donner lieu à quelques débats, si bien qu'il nous faut préciser ce que nous entendons par « identité de dénotation » ou « coréférence ».

Mis à part le problème philosophique que peut représenter le fait de désigner des êtres qui n'existent pas, problème que notre définition de la dénotation écarte (voir la section 1.4), le problème qui peut se poser pour la définition de la coréférence est celui des « référents évolutifs ». Nous reprenons la présentation que M. Charolles et C. Schnedecker [19] font de ce problème. Ces deux auteurs citent le texte suivant :

(1) « En 1908, *un petit prince de trois ans* est enlevé à sa mère et assis sur le trône laqué de l'Empire de Chine. Pendant 16 ans, *il* reste un demi-dieu prisonnier de la Cité interdite. Puis *il* mène une vie de play-boy insouciant sous protection japonaise, se retrouve empereur de Mandchourie, est arrêté par les Russes et rendu à quarante-trois ans aux Chinois qui *le*

[1] Avec quelques restrictions : dans MUC, les expressions en fonction d'attibut peuvent entrer dans une relation de coréférence, alors que, dans notre système, elles ne sont pas dénotantes et ne peuvent donc être interprétées avec une identité de dénotation.

> rééduquent dans un camp pendant dix ans » (présentation du film « Le dernier empereur », magazine de télévision)

et remarquent qu'« on peut se demander si les différents pronoms qui apparaissent dans le cours du récit renvoient encore bien au *petit prince de trois ans*, car on imagine mal que celui-ci puisse littéralement mener une vie de play-boy ou être rééduqué par les Chinois. »

La principale raison pour laquelle on met en avant de tels exemples est de montrer qu'on n'interprète pas les pronoms en les remplaçant littéralement par leur antécédent. Ces exemples peuvent cependant induire pour certains une mise en cause de la notion de coréférence. Le commentaire suivant, d'A. Reboul [76], va dans ce sens, dans la mesure où l'auteur semble mettre en cause non seulement le fait qu'on puisse substituer l'antécédent au pronom, mais aussi le fait que les deux expressions « désignent le même objet » :

> Les exemples de référents évolutifs [sont] extrêmement importants, dans la mesure où ils contredisent une thèse généralement admise sur l'interprétation des pronoms, la *thèse subtitutionaliste*. Selon cette thèse, un pronom est interprété à partir de son antécédent, désigne le même objet et on peut lui substituer l'antécédent en question, à partir duquel on peut lui attribuer le « bon » référent.

Pour notre part, il ne fera aucun doute que les expressions en italiques dans le texte (1) ci-dessus sont coréférentes entre elles. Le passage suivant, de Frege [34, p. 162], nous permettra d'expliciter notre position sur la question de l'identité de dénotation :

> Quand quelque chose varie, différentes propriétés et états affectent successivement le même objet. S'il n'était pas le même, il n'y aurait aucun sujet dont on puisse énoncer qu'il varie. Un bâton s'allonge sous l'effet de la chaleur ; pendant l'échauffement, il demeure le même. Si au contraire on l'avait emporté et remplacé par un autre bâton plus long, on ne pourrait pas dire qu'il s'est allongé. Un homme prend de l'âge, mais si nous ne pouvions pas reconnaître en lui le même homme, il n'y aurait rien dont nous puissions énoncer l'âge.

Nous dirons que deux expressions sont coréférentes, ou qu'il y a identité de dénotation entre deux expressions, si elles désignent le même être de l'univers de dénotation au-delà des variations qu'il est susceptible de subir dans le temps.

Cette définition étant posée, nous nous attachons, dans les sections qui suivent, à décrire différents cas de reprise avec identité de dénotation.

2.1.2 Cas général

La relation de coréférence met en jeu les êtres dénotés par les expressions. Pour noter l'observation des relations de coréférence dans un texte, on représente

les êtres de l'univers de dénotation associé au texte par des « index de référents » de la forme $\mathtt{o}_i$ (pour « objet i »), où i tient la place d'un nombre entier positif. Pour distinguer deux êtres distincts, on utilise deux entiers distincts. Si deux expressions distinctes dénotent le même être (c'est le cas lorsque la relation de reprise est une identité de dénotation), on utilise le même index de référent pour chacune des expressions.

Pour décrire un lien de reprise, l'observateur doit placer les expressions source et reprise entre crochets angulaires et noter entre crochets carrés les informations nécessaires à la description du type de reprise (voir la section 1.5). Lorsque la relation de reprise est une identité de dénotation, l'observateur note simplement à la suite de chacune des expressions l'index de référent attribué à l'être que ces deux expressions dénotent.

L'annotation de la phrase suivante,

(2) <La BNP [o1]> a annoncé jeudi la division de < <ses [o1]> activités [o2]> de marché, <dont [o2]> l'importance va croissante, en deux pôles séparés.

signifie que le déterminant possessif *ses* est interprété comme ayant la même dénotation que l'expression *la BNP* et que le pronom relatif *dont* est interprété comme ayant la même dénotation que le syntagme *ses activités de marché* [2].

L'annotation du texte suivant :

(3) <Robert Panhard [o1]>, cinquante-deux ans, a été élu hier président de la Chambre des notaires de Paris. Titulaire d'une maîtrise de droit et diplômé de l'Institut des études politiques de Paris, <il [o1]> débute <sa [o1]> carrière comme fondé de pouvoir à la BIMP (Banque Industrielle et Mobilière Privée). Diplômé du notariat en 1979, <il [o1]> rejoint la société civile professionnelle Dauchez, Kubisa, Panhard, Baffoy et Deneuville. <Il [o1]> entend désormais valoriser la profession, en constituant un réseau notarial européen, ancré à Paris, renforcer le pôle immobilier et moderniser les études notariales.

signifie que l'ensemble des expressions auxquelles a été associé l'index de référent o1 sont interprétées comme désignant toutes le même être de l'univers de dénotation : la personne qui s'appelle Robert Panhard.

Dans le texte suivant,

(4) <La Société Générale [o1]> ne digère toujours pas la décision du gouvernement de <céder le CIC au Crédit Mutuel [o2]>. Dans une interview publiée ce week-end par le Journal des finances, Daniel Bouton, le président de <la banque [o1]> privée, continue de faire planer la menace d'une action visant à remettre en question <l'opération [o2]>.

[2] On rappelle que, pour les syntagmes nominaux, la convention est de ne placer entre crochets angulaires que le syntagme nominal noyau (voir la section 1.5.1).

on considère que l'expression *la banque* doit être interprétée comme ayant la même dénotation que *La Société Générale*. On considère également que l'expression *l'opération* fait référence à la cession du CIC au Crédit Mutuel par le gouvernement, donc qu'elle a la même dénotation que la proposition *céder le CIC au Crédit Mutuel* [3]. On le voit, la source d'une reprise avec identité de dénotation peut être une proposition ou une phrase aussi bien qu'un syntagme nominal.

Autre exemple dans lequel un syntagme nominal défini a pour source une proposition, avec laquelle il est coréférent :

(5) Après avoir cédé sa banque d'affaires à la Société Générale, <le groupe britannique Hambros vient de vendre sa participation de 52 % dans Hambro Insurance Services Group au canadien Lindsey Morden Groupe `[o1]`>. Le montant de <la transaction `[o1]`> est de 44,9 millions de livres (environ 436 millions de francs), soit 132 pence par action Hambro Insurance Services, de quoi faire ressortir une prime de 5,2 % par rapport au cours moyen du 7 mai dernier.

Enfin, l'identité de dénotation peut aussi être observée entre deux phrases, comme en témoigne l'exemple suivant, repris de [27] :

(6) <Fred a abîmé un vêtement `[o1]`>. <Il a taché une chemise `[o1]`>.

Les deux phrases désignent ici le même événement.

2.1.3 Désignation d'ensembles

Nous avons distingué (section 1.4.4) dans l'univers de dénotation des êtres qui sont des êtres singuliers et des êtres qui sont des ensembles d'au moins deux êtres singuliers. Certains liens de reprise avec identité de dénotation mettant en jeu de tels êtres nécessitent une notation particulière.

Reprise dont la source est un ensemble d'expressions

Il est possible qu'une expression e_i qui dénote un ensemble d'êtres singuliers s'interprète non pas en fonction d'une expression qui la précède, mais en fonction de plusieurs expressions, telles que l'ensemble ou l'union des êtres dénotés par chacune de ces expressions spécifie la dénotation de l'expression e_i. Dans le texte suivant,

(7) La pression monte chaque jour d'un cran à Milan pour pousser à un mariage entre la Comit et Banca di Roma, qui donnerait naissance à la première banque italienne. Les informations s'accumulent, laissant penser que le processus visant à préparer la fusion est lancé. Ainsi, les deux banques ont déjà choisi leurs banques conseils (Merrill Lynch pour la Comit et Goldman Sachs pour Banca di Roma).

[3] Étant entendu que, dans la phrase où elle apparaît, cette proposition a pour sujet implicite *le gouvernement*.

l'expression *les deux banques* dénote l'ensemble constitué des êtres singuliers que sont la Comit et Banca di Roma, dénotés dans la première phrase par deux expressions distinctes. C'est en fonction de ces deux expressions que l'expression *les deux banques* est interprétée.

Nous décrivons de tels cas de figure, où l'extension d'un ensemble dénoté par une expression e_i est spécifiée par plusieurs expressions e_j, ..., e_n qui précèdent et dénotent chacune un élément — ou éventuellement un sous-ensemble — de e_i, comme des reprises avec identité de dénotation. Dans cette optique, la coréférence est alors une relation entre une expression et un ensemble d'expressions.

La notation qui sera utilisée pour ces cas est la suivante. Chacune des expressions entrant en relation est annotée avec un index de référent, et, à la suite de la phrase contenant le second terme de la reprise, on place une formule de la forme `[{`o_j, ..., o_n`}-id-`o_i`]`, où o_j, ..., o_n sont les index de référent associés aux expressions sources et o_i l'index associé à la reprise. Pour l'exemple ci-dessus, cela donne :

(8) La pression monte chaque jour d'un cran à Milan pour pousser à un mariage entre <la Comit `[o1]`> et <Banca di Roma `[o2]`>, qui donnerait naissance à la première banque italienne. Les informations s'accumulent, laissant penser que le processus visant à préparer la fusion est lancé. Ainsi, <les deux banques `[o3]`> ont déjà choisi leurs banques conseils (Merrill Lynch pour la Comit et Goldman Sachs pour Banca di Roma).
<`[{o1,o2}-id-o3]`>

Dénotation générique

Nous avons également distingué dans la section 1.4.4 certains cas où une expression dénote une classe d'êtres de l'univers de dénotation, situations que nous décrivons comme des cas de « dénotation générique ». L'observateur doit distinguer de tels usages par l'ajout du symbole `(G)` à la suite de l'index de référent attribué à l'être (c'est-à-dire la classe) dénoté par le syntagme. Si une expression référentielle à valeur générique est reprise par une expression qui a la même dénotation, il n'est pas nécessaire de noter à nouveau le symbole `(G)` au niveau de la reprise : l'identité d'index, d'une part, et la mention de l'emploi générique sur la première expression, d'autre part, suffisent.

Le texte suivant offre un exemple de distinction entre un référent spécifique et un référent générique. Il est fait référence à un gorille particulier (index de référent `o1`), mais aussi à la classe des gorilles (index `o2`). Seules les reprises qui concernent l'identité de ces deux référents sont marquées.

(9) C'est à travers de larges grilles, que les femelles du canton contemplaient <un puissant gorille `[o1]`>, sans souci du qu'en dira-t-on ; [...] Tout à coup, la prison bien close, où vivait <le bel animal `[o1]`>, s'ouvre, on ne sait pourquoi (je suppose qu'on avait dû la fermer mal) ; [...] Dès que

> la féminine engeance sut que <le singe [o1]> était puceau, au lieu de profiter de la chance, elle fit feu des deux fuseaux ! Celles-là même qui, naguère, <le [o1]> couvaient d'un oeil décidé, fuirent, prouvant qu'elles n'avaient guère de la suite dans les idées ; d'autant plus vaines étaient leur crainte, que <le gorille [o2(G)]> est un luron supérieur à l'homme dans l'étreinte, bien des femmes vous le diront !

Dans le texte suivant, le pronom *il* fait référence, comme le syntagme nominal *Le lion*, à tout membre potentiel de la classe des lions. On ne marque que l'être désigné est une classe qu'au niveau de la source ; l'identité de l'index de référent (o1) suffit à dire que *il* est aussi employé dans un sens générique.

(10) <Le lion [o1(G)]> est un grand chasseur ; <il [o1]> ne rate jamais sa proie.

Modalités particulières dans les désignations d'ensembles.

On distingue, parmi les reprises avec identité de dénotation mettant en jeu une désignation d'un ensemble, deux modalités particulières : la désignation avec valeur distributive, liée à l'emploi des formes *chaque* et *chacun*, et la désignation avec valeur négative, liée à l'emploi des pronoms *aucun* ou *nul*.

VALEUR DISTRIBUTIVE. Voici ce que Grevisse dit des formes *chaque* et *chacun* [37, §611 et §717] :

> *Chaque*, qui ne s'emploie qu'au singulier, est un déterminant distributif, c'est-à-dire qu'il marque que l'on considère en particulier les divers éléments d'un ensemble.
>
> *Chacun* [...] a une valeur distributive, c'est-à-dire qu'il s'emploie quand on considère un à un, isolément, les éléments d'un ensemble.

Nous considérerons donc que les formes *chaque* et *chacun* désignent un ensemble. Si une expression avec l'une ou l'autre de ces formes est interprétée par reprise, nous considérons que cette reprise relève de la coréférence. Pour distinguer la valeur distributive de l'expression, on utilise simplement la notation habituelle, à cette différence près qu'on complète avec le symbole (D) l'index de référent de l'expression à valeur distributive.

Quelques exemples.

(11) <Les deux groupes [o1]> cherchent à formaliser un pacte d'actionnaires stipulant les compétences de <chacun [o1(D)]> dans le cadre de la réforme des Caisses d'Épargne.

(12) <Les magistrats [o1]> doivent avoir <chacun [o1(D)]> un bureau spacieux pour les auditions.

(13) Pierre reçoit <ses amis [o1]>. <Chaque invité [o1(D)]> a apporté quelque chose.

S'il y a reprise de l'expression à valeur distributive, le symbole `(D)` est maintenu au niveau de la reprise.

(14) <Pierre `[o1]`> et <Marie `[o2]`> jouent <chacun `[o3(D)]`> de <leur `[o3(D)]`> côté.
<`[{o1,o2}-id-o3]`>

(15) <Pierre `[o1]`> et <Marie `[o2]`> jouent <chacun `[o3(D)]`> de <son `[o3(D)]`> côté.
<`[{o1,o2}-id-o3]`>

Valeur négative. Les formes *aucun* et *nul* ont le plus souvent une valeur négative. On traite cependant les syntagmes nominaux contenant ces formes comme des expressions dénotantes avec valeur distributive — c'est-à-dire, s'ils sont des reprises, comme des reprises avec identité de dénotation — et transfert de la négation sur le verbe dont dépend le syntagme [4]. Ce verbe est le plus souvent dans le langage écrit précédé par la particule *ne*.

La phrase *nul n'en est revenu* dans le discours suivant :

(16) Plusieurs explorateurs sont allés dans ces régions ; nul n'en est revenu.

est interprétée par nous comme « quel que soit l'être appartenant à l'ensemble dénoté par *Plusieurs explorateurs*, cet être n'en est pas revenu. » Avec `o1` l'ensemble dénoté par *Plusieurs explorateurs*, on a la glose « chacun des membres de `o1` n'en est pas revenu. » On utilise donc la notation prévue pour la distribution :

(16) <Plusieurs explorateurs`[o1]`> sont allés dans ces régions ; <nul `[o1(D)]`> n'en est revenu.

Le fait qu'on utilise la même notation ne doit pas être interprété comme signifiant strictement que *chacun* et *aucun* (ou *nul*) sont équivalents, mais plutôt qu'il y a une proximité entre ces termes, qui est qu'ils sont interprétés relativement à un ensemble. La notation $\mathtt{o}_i\mathtt{(D)}$ fait abstraction de la différence entre ces formes, différence qui, si elle n'apparaît pas dans la notation, doit être néanmoins prise en compte pour l'interprétation complète de la phrase.

Dans l'exemple suivant, nous interprétons la phrase *Aucun syndicat n'envisage vraiment de...* comme « chacun des membres de l'ensemble `o1` n'envisage pas vraiment de... ».

(17) <Les syndicats `[o1]`> du secteur bancaire s'apprêtent à reprendre le chemin des négociations de branche d'ici fin mai. [...] <Aucun syndicat `[o1(D)]`> n'envisage vraiment de rester en dehors des négociations.

Dans le texte suivant, il est dit que quel que soit l'être appartenant à l'ensemble constitué des êtres Pierre, Jacques et Marie, cet être ne retournera pas en Provence :

[4] Étant donné une phrase de la forme *aucun x n'est P* ($\neg\exists x,\ P(x)$), on passe à la formulation équivalente *quel que soit, x n'est pas P* ($\forall x,\ \neg P(x)$).

(18) <Pierre [o1]>, <Jacques [o2]> et <Marie [o3]> ont passé <leurs [o4]> vacances en Provence. <Aucun [o4(D)]> n'y retournera.
<[{o1,o2,o3}-id-o4]>

Si un syntagme nominal avec *aucun* ou *nul* est source d'une reprise sans être lui-même une reprise, la notation ne change pas. On a là quelque chose qui s'apparente à une distribution sur un ensemble générique, une classe d'êtres de l'univers de dénotation.

(19) <Aucun groupe [o1(D)]> étranger n'a pour le moment fait connaître <son [o1(D)]> intérêt.

Glose : « Parmi tous les groupes étrangers, chacun des groupes n'a pas fait connaître son intérêt. »

Avec cette description des modalités particulières dans les références aux ensembles, nous terminons la description des liens de reprise avec identité de dénotation.

2.2 Identité de description

Les liens de reprises que nous avons observés dans la section précédente mettent en jeu une identité de dénotation. Nous décrivons dans cette section des liens de reprises dont nous disons qu'ils mettent en jeu une identité de description. On distingue deux principaux cas de figure selon qu'il y a :

1. reprise par une expression dénotante,
2. reprise par une expression non dénotante.

Nous nous limitons dans cette section essentiellement à la description de liens de reprise avec identité de description qui mettent en jeu une expression anaphorique [5]. Les reprises avec identité de description (dites aussi « reprises de description ») seront caractérisées au travers d'exemples, qui nous permettrons de dresser la liste des expressions qui sont le plus souvent interprétées de cette manière. On notera que là où l'identité de dénotation pouvait mettre en jeu un nombre quelconque d'expressions coréférentes, l'identité de description, lorsqu'elle met en jeu une expression anaphorique, est en général une relation entre deux expressions.

Le principe de notation pour les identités de description sera le même que pour les identités de dénotation : délimitation des expressions par des chevrons

[5] On entend par « expressions anaphoriques » des expréssions qui appartiennent à l'une des catégories répertoriées page 23. Dans l'expérience qui sera décrite au chapitre 4, seuls ces liens devaient être annotés pour les identités de description. Les identités de description qui ne mettent pas en jeu une expression anaphorique sont en fait des répétitions de description et il nous a semblé que leur observation serait assez triviale (voir l'exemple (14) du chapitre 1, page 29).

et informations décrivant le lien de reprise entre crochets. On représentera la description en jeu dans le lien de reprise par une chaîne de caractères correspondant au lemme du noyau de l'expression source, chaîne de caractère qui remplira le rôle des index de référent dans les reprises dénotationnelles. Des exemples seront présentés dans les sections qui suivent.

2.2.1 Reprise par une expression dénotante

Une expression dénotante peut être interprétée par reprise de description. La source est, en règle générale, elle aussi une expression dénotante. Ces expressions dénotent alors des êtres différents mais qui peuvent être décrits de la même manière, autrement dit des êtres qui sont du même type.

Pronoms démonstratifs simples

Un cas typique de reprise de description par une expression dénotante se rencontre avec les formes simples du pronom démonstratif (*celui, celle, ceux, celles*). Dans l'exemple suivant, pour identifier l'être désigné par le syntagme *celles de la Méditerranée*, il faut lui associer la description *plages*, reprise du syntagme *les plages de l'Atlantique*. On note entre crochets le lemme *plage* au niveau de la source et de la reprise.

(20) Marie aime <les plages `[plage]`> de l'Atlantique ; Jeanne préfère <celles `[plage]`> de la Méditerranée.

On remarquera que, bien que seule la description nominale soit reprise ici, on met entre crochets angulaires le syntagme nominal noyau complet. Cette notation est adoptée pour limiter le parenthésage des expressions. Chacune des expressions *les plages de l'Atlantique* et *celles de la Méditerrannée* est une expression dénotante et le fait qu'elle soit la source d'une reprise avec identité de dénotation est envisageable. S'il est nécessaire par exemple d'attribuer des index de référent au référent des expressions *les plages de l'Atlantique* et *celles de la Méditerranée*, on note ces index à la suite de la chaîne de caractères `plage` :

(21) Marie aime <les plages `[plage, o1]`> de l'Atlantique ; Jeanne préfère <celles `[plage, o2]`> de la Méditerranée. C'est un fait que <celles-ci `[o2]`> sont plus jolies que <celles-là `[o1]`>.

Syntagmes à noyau adjectival

Quand on désigne deux objets différents mais de même type dans un segment de discours réduit, on peut se dispenser de mentionner au niveau de la seconde expression la description qui leur est commune, soit au moyen du pronom démonstratif, comme ci-dessus, soit, si l'expression contient un adjectif, en ne répétant

pas la description nominale. On a alors un syntagme nominal à noyau adjectival, situation qui peut aussi être décrite comme une ellipse du noyau nominal du syntagme [6]. Dans la phrase suivante, le syntagme *les blancs* doit être interprété comme *les chevaux blancs*. La description *cheval* est reprise du syntagme *un cheval noir*.

(22) Il avait <un cheval `[cheval]`> noir. Elle préférait <les blancs `[cheval]`>.

On notera que dans la mesure où le nombre varie entre les deux syntagmes (singulier pour la source et pluriel pour l'expression anaphorique), il convient lorsqu'on parle de description de faire abstraction des variations morphologiques des signifiants.

Pronoms possessifs

Les « pronoms possessifs » (*le mien, la sienne, les leurs...*) peuvent être analysés de la même manière que les syntagmes à noyau adjectival, si on les considère comme une séquence déterminant défini + adjectif. Une telle analyse permet de mettre à jour la présence d'un double lien anaphorique : identité de description au niveau du syntagme complet (par exemple *la sienne*) et identité de dénotation au niveau de l'adjectif (par exemple *sienne*). Dans le texte suivant, le syntagme *la sienne* dénote un être de type *fille*, ce qu'on interprète par reprise de la description qui est faite de l'être dénoté par *la fille de Jean* :

(23) Marie aime <la fille `[fille]`> de Jean. <Léa `[o1]`> préfère <la <sienne `[o1]`> `[fille]`>.

L'annotation de *sienne* par l'index de référent `o1` indique que l'adjectif *sienne* est interprété comme ayant la même dénotation que l'expression *Léa*, c'est-à-dire qu'on interprète le syntagme *la sienne* comme dénotant la fille de Léa.

Identité vague

Le double lien de reprise qu'on rencontre avec les pronoms possessifs se retrouve dans les cas d'*identité vague* [7]. Une expression pronominale, le plus souvent un déterminant possessif, apparaissant dans la source d'une reprise de description est réinterprétée au lieu d'être reprise avec sa dénotation initiale. Une des lectures de la phrase suivante veut que Jeanne dépense son propre salaire et non celui de Marie ou d'une tierce personne. Dans cette lecture, il y a reprise de la description *son salaire* et réinterprétation du possessif comme coréférent avec *Jeanne*.

(24) <Marie `[o1]`> dépose < <son `[o1]`> salaire `[salaire]`> à la banque et <Jeanne `[o2]`> <le `[<son [o2]> salaire]`> dépense aussitôt.

[6] Plutôt que d'annoter un segment vide à l'emplacement de la tête nominale manquante, nous préférons annoter l'ensemble du syntagme nominal à noyau adjectival, c'est-à-dire l'expression dénotante.

[7] Notre traduction du terme anglais *sloppy identity*. Voir à ce sujet T. Reinhart [78, p. 62].

2.2.2 Reprise par une expression non dénotante

Certaines expressions anaphoriques sont non dénotantes et interprétées par reprise de description.

Pronom clitique objet d'un verbe copule.

Dans notre terminologie, une expression en fonction d'attribut est considérée comme non dénotante (voir section 1.4.3). Il s'ensuit que le clitique *le* objet d'un verbe à valeur de copule (*être, sembler, rester, demeurer, paraître*, etc.) sera non dénotant et interprété par reprise de description.

Dans l'exemple suivant, le clitique *l'* reprend la description qui est faite de l'objet dénoté par *Marie* au moyen de l'expression *belle*, et seulement cette description. Dans le cas où la description est effectuée avec un verbe copule, on ne place pas le verbe dans les crochets angulaires. On place donc entre crochets angulaires la seule description *belle* et le clitique *l'* et on leur associe le lemme `beau`.

(25) Marie est <belle `[beau]`>. Juliette ne <l' `[beau]`> est pas.

Le texte suivant est un autre exemple de reprise ayant pour source une description avec un verbe copule :

(26) Mais voici que se profile la deuxième cohorte : celle qui trouve que la technologie, loin d'être trop <performante `[performant]`>, ne <l' `[performant]`> est pas assez.

La description qui est reprise peut être exprimée par le noyau d'une expression dénotante :

(27) <Le Sage `[sage]`>, autre surnom du président ne <l' `[sage]`> était plus depuis longtemps.

(28) Le document avait été présenté par le ministre comme étant le moyen de faire en sorte que <les pauvres `[pauvre]`> <le `[pauvre]`> soient de moins en moins.

La source peut être une épithète :

(29) Les restructurations effectuées dans cette branche <déficitaire `[déficitaire]`> (la seule qui <le `[déficitaire]`> soit chez Siemens avec l'informatique) auraient permis de gros progrès.

ou un participe passé modifiant un syntagme nominal à droite :

(30) Sur un total de 2 494 400 chômeurs < indemnisés `[indemniser]` >, 2 092 300 <l' `[indemniser]`> ont été au titre du régime d'assurance stricto sensu.

(31) Le traitement le plus favorable <accordé `[accorder]`> à un pays membre du GATT doit <l' `[accorder]`> être à tous les autres.

ou un verbe à la voix passive :

(32) Si Jacques a été <licencié `[licencier]`>, Pierre <le `[licencier]`> sera aussi.

Ellipse verbale

Nous parlons d'ellipse verbale lorsque le verbe d'une proposition n'est pas exprimé et que celui-ci doit être retrouvé grâce au contexte textuel proche. Pour noter une ellipse verbale, on met à l'emplacement de l'ellipse une paire de crochets angulaires contenant, entre crochets carrés, le lemme associé à la description qui est reprise. Comme pour le clitique objet d'un verbe copule, si la description source est exprimée par un adjectif ou participe passé construit avec un verbe copule, on n'annote que l'adjectif ou le participe passé.

(33) Jacques a été <mordu `[mordre]`> par un chien, Pierre <`[mordre]`> par un chat.

L'exemple suivant met en lumière la proximité d'interprétation du clitique *le* avec verbe copule et de l'ellipse verbale.

(34) L'accord de libre-échange nord-américain signé le 7 octobre 1992 par le Mexique, le Canada et les États-Unis est <ratifié `[ratifier]`> par le Congrès américain le 20, et <`[ratifier]`> par le Sénat mexicain le 22, après <l' `[ratifier]`> avoir été par le Parlement canadien en mai 1993.

Dans ce type de reprise, l'ellipse verbale est en générale repérable par la présence d'expressions qui jouent le rôle de compléments ou modifieurs du verbe non exprimé. Dans l'exemple (33), le syntagme *par un chat* est complément d'agent du verbe non exprimé. Dans l'exemple (34) les syntagmes *par le Sénat mexicain* et *le 22* sont respectivement complément d'agent et complément circonstanciel de temps du verbe non exprimé.

Certains adverbes, tels que *aussi* et la particule de négation *pas* dans les deux exemples suivants indiquent plus nettement la présence éventuelle d'une ellipse verbale.

(35) Marie <mange une pomme `[manger]`> ; Juliette <`[manger]`> aussi.

(36) Marie <dort dans sa chambre `[dormir]`>. Juliette <`[dormir]`> pas.

Reprise des expressions dénotantes associées à la description source

Dans les exemples de reprises de description par une expression non dénotante présentés jusqu'à présent, la source était toujours une simple description. La source peut aussi être plus complexe. Dans les exemples ci-dessous, en même temps que la description verbale ou adjectivale elle-même, l'expression anaphorique reprend une ou plusieurs expressions dénotantes associées à la description qui est reprise. Il y a implicitement reprise avec identité de dénotation pour ces

expressions. On met entre crochets angulaires l'ensemble constitué par la description qui est reprise et les expressions dénotantes. On ne met entre crochets carrés que le lemme associé à la description, étant entendu que ce qui est repris est tout ce qui est entre les crochets angulaires [8].

Quelques exemples, suivis à chaque fois d'une glose explicitant l'interprétation qu'on donne à la phrase.

(37) Si Blair House 1 n'était pas <pour nous acceptable `[accepter]`>, a-t-il ajouté, Blair House 2 ne <l' `[accepter]`> est pas davantage.

Glose : « Blair House 2 n'est pas davantage *acceptable pour nous* ». La dénotation de *nous* est reprise en même temps que la description *acceptable*. Dans la suite *pour nous acceptable*, l'adjectif *acceptable* (expression non dénotante) est le noyau, dont dépend le syntagme *pour nous* (expression dénotante). Outre le caractère non dénotant de l'expression anaphorique, c'est le caractère non dénotant du noyau de l'expression source qui fait des reprises présentées dans cette sous-section des reprises avec identité de description, plutôt qu'avec identité de dénotation.

(38) Désormais, les décisions <prises par les deux sociétés `[prendre]`> <le `[prendre]`> seront sur une base paritaire.

Glose : « les décisions prises par les deux sociétés seront *prises par les deux sociétés* sur une base paritaire ». La dénotation de *les deux sociétés* est reprise en même temps que la description *prendre*.

(39) Ce que <la France a obtenu de l'Allemagne `[obtenir]`> sur l'agriculture ne peut <l' `[obtenir]`> être sur l'audiovisuel.

Glose : « Ce que la France a obtenu de l'Allemagne sur l'agriculture ne peut être *obtenu de l'Allemagne par la France* sur l'audiovisuel ». La dénotation de *la France* et la dénotation de *l'Allemagne* sont reprises en même temps que la description *obtenir*.

(40) <Volvo détient `[détenir]`> actuellement 35 % de RVA, mais 17,15 % <le `[détenir]`> sont indirectement, via le holding RVC.

Glose : « Volvo détient actuellement 35 % de RVA, mais 17,15 % sont *détenus par Volvo* indirectement ». La dénotation de *Volvo* est reprise en même temps que la description *détenir*.

Une ellipse verbale peut aussi reprendre les expressions dénotantes associées à une description :

(41) Marie <aime aider les pauvres `[aimer]`>, Jacques <`[aimer]`> pas.

Le plus souvent, la source est constituée d'une suite ininterrompue d'expressions ; on place cette suite entre crochets angulaires et on n'utilise que le lemme

[8] Sauf, éventuellement, si on a affaire à un cas d'identité vague (voir p. 51).

verbal pour indiquer le lien de reprise. S'il y a à l'intérieur des crochets des éléments descriptifs qui ne sont pas repris, on délimite ces éléments par des symboles spéciaux : `/-` *élément non repris* `-/`.

(42) « Une solution qui vaut mieux qu'une victoire écrasante de la Grèce », affirme-t-il, reconnaissant qu'une telle position aurait été <considérée `/-`, il y a peu de temps encore, `-/` comme une « hérésie » `[considérer]`>. Elle <l' `[considérer]`> est d'ailleurs encore aux yeux de certains membres de la majorité.

Glose : « elle est d'ailleurs encore *considérée comme une hérésie* aux yeux de certains membres de la majorité ». Le segment *il y a peu de temps encore* n'est pas repris.

Si la source à noyau verbal d'une reprise de description contient une expression pronominale, celle-ci pourra donner lieu à un cas d'*identité vague* (voir page 51). Dans ce cas, on explicitera entre les crochets au niveau de l'expression anaphorique l'ensemble des expressions qui sont reprises et on notera la nouvelle interprétation de l'expression pronominale [9].

(43) <Marie `[o1]`> <aime promener <son `[o1]`> chien `[aimer]`> ; <Jacques `[o2]`> <`[aimer promener <son [o2]> chien]`> pas.

Pro-verbe *le faire*

On analyse une séquence clitique *le* plus verbe *faire* comme une reprise de description :

(44) Marie <n'a pas promené le chien `[promener]`>, puisque Jeanne voulait <le faire `[promener]`>.

Lorsque le syntagme verbal source ou reprise contient des particules de négation ou des adverbes, on les inclut dans l'expression délimitée, étant entendu que si ils sont présents au niveau de la source, ces éléments ne sont pas repris :

(45) Marie <n'a pas promené le chien `[promener]`>, puisque Jeanne <l'a fait `[promener]`>.

Reprise de description avec le clitique *en*

Le clitique *en* peut dans certains cas être non dénotant et être interprété comme une reprise de description.

Dans l'exemple suivant, les expressions *trois enfants* et *deux* dénotent deux êtres différents, mais ayant en commun d'être décrits comme des *enfants*. Il y a identité de description entre *deux* et *trois enfants*. On considère que le clitique

[9] L'annotation de cet exemple correspond à l'interprétation selon laquelle c'est son propre chien que Jacques n'aime pas promener et non celui de Marie.

en est non dénotant et on l'interprète lui aussi comme une reprise avec identité de description.

(46) Pierre et Marie ont <trois enfants `[enfant]`> ; Jacques et Juliette <en `[enfant]`> ont <deux `[enfant]`>.

Autre exemple, dans lequel il n'y a pas, outre la reprise de description avec *en*, de reprise par une expression dénotante :

(47) Pierre et Marie ont <trois enfants `[enfant]`> ; Jacques et Juliette <en `[enfant]`> ont aussi.

Ces deux exemples contrastent avec le suivant, dans lequel nous interprétons le clitique *en* comme une reprise avec identité de dénotation :

(48) Pierre et Marie ont <trois enfants `[o1]`>, mais Jacques n'<en `[o1]`> connaît que <deux `[o2]`>.

Ce que dit la seconde proposition de cette phrase est que Jacques ne connaît que deux des trois enfants de Pierre et Marie. On interprète le syntagme *deux* comme dénotant un sous-ensemble de l'être dénoté par *trois enfants* et on interprète le pronom *en* comme dénotant également les trois enfants de Pierre et Marie [10].

Noms à valeur de numéral

Les syntagmes nominaux ayant pour noyau un nom à valeur de numéral (p. ex.*million, vingtaine*) sont en général accompagnés d'un syntagme prépositionnel complément dont le noyau décrit l'être désigné par le syntagme. Ce complément peut cependant être omis, et il peut alors être nécessaire de reprendre une description explicite dans le contexte :

(49) Les deux filiales du Crédit Lyonnais ont affiché en 1997 des pertes de <4,3 milliards `[peseta]`> de pesetas pour le CL-Espana et de <2,6 milliards `[peseta]`> pour Banca Jover.

Dans cet exemple, on considère que les expressions *4,3 milliards de pesetas* et *2,6 milliards* sont non dénotantes ; elles ne font qu'exprimer la mesure des pertes des sociétés en question. On remarquera qu'on note toujours la description au niveau du syntagme nominal noyau. Par ailleurs, on notera que ces mêmes expressions pourront, dans un autre contexte, être des expressions dénotantes (comme, par exemple, dans la phrase *4,3 milliards de pesetas ont été perdus par la société, qui devait les utiliser pour son projet de développement.*).

2.3 Relation « membre-de »

Nous avons vu que deux expressions peuvent être reliées par une identité de dénotation ou une identité de description. Nous présentons dans cette section une

[10] La notation prévue pour le cas illustré en (48) sera présentée dans la section 2.3.

relation qui met en jeu à la fois une identité de dénotation partielle et une identité de description. Cette relation est celle qui lie les éléments ou les sous-ensembles d'un ensemble à l'ensemble dont ils sont éléments ou dans lequel ils sont inclus. Nous appelons cette relation « `membre-de` ». Comme la coréférence, la relation `membre-de` concerne la dénotation d'expressions dénotantes, mais nous verrons qu'elle met aussi en jeu une identité de description.

Nous définirons la relation `membre-de` que nous utilisons pour rendre compte des phénomènes de reprise en la mettant en rapport avec les notions d'appartenance et d'inclusion de la théorie des ensembles. Il est cependant capital de bien distinguer que l'expression *x est membre de y* dans le langage que nous utilisons pour décrire les reprises n'est pas synonyme de l'expression *x est membre de y* que l'on utilise parfois en théorie des ensembles. De fait, la relation que nous utilisons recouvre à la fois les notions d'appartenance et d'inclusion de la théorie des ensembles. Elle recouvre également la relation entre une classe et une instance ou un ensemble d'instances de cette classe.

2.3.1 Exemples

Quelques exemples, avant d'en venir à une définition plus précise de la relation `membre-de`. Pour noter qu'un être $\mathtt{o}_i$, dénoté par une expression e_i, est membre d'un être $\mathtt{o}_j$, dénoté par une expression e_j, nous placerons à la suite des expressions e_i et e_j l'index de référent correspondant, et nous utiliserons une formule de la forme $\mathtt{o}_i$`-mde-`$\mathtt{o}_j$ pour indiquer la relation entre les deux référents. Cette formule se lit « l'être $\mathtt{o}_i$ est membre de l'être $\mathtt{o}_j$ ». On la place entre chevrons et crochets, de préférence à la suite de la phrase qui contient l'expression qui apparaît en deuxième dans le texte.

Dans la phrase suivante, nous dirons que l'être dénoté par *la plupart* est `membre de` l'être dénoté par *les manifestants*.

(50) <Les manifestants `[o1]`> se sont dispersés, mais <la plupart `[o2]`> sont restées en ville.
<`[o2-mde-o1]`>

Dans l'exemple suivant, nous dirons que l'être dénoté par *le blanc* est `membre` de l'être dénoté par *deux chevaux*. Le cheval pris par Marie est l'un des deux chevaux de Pierre.

(51) Pierre avait <deux chevaux `[o1]`>. Marie a pris <le blanc `[o2]`>.
<`[o2-mde-o1]`>

Dans l'exemple suivant, nous dirons que l'être dénoté par *celui qu'elle préférait* est `membre de` l'être dénoté par *deux chevaux*.

(52) Pierre avait <deux chevaux `[o1]`>. Marie a pris <celui `[o2]`> qu'elle préférait.
<`[o2-mde-o1]`>

Dans l'exemple suivant, nous interprétons le clitique *en* comme coréférent avec *trois enfants* et nous dirons que l'être dénoté par *deux* est `membre de` l'être dénoté par ces deux premières expressions. Ce sont deux des enfants de Pierre et Marie que Jacques connaît.

(53) Pierre et Marie ont <trois enfants `[o1]`> ; Jacques <en `[o1]`> connaît <deux `[o2]`>.
<`[o2-mde-o1]`>

Nous utilisons aussi la relation `membre-de` pour décrire la relation entre une classe et une instance ou un ensemble d'instances de cette classe. Étant donné la phrase suivante,

(54) <Le lion `[o1(G)]`> est peut-être un grand chasseur, mais <celui `[o2]`> que Pierre a tué ce matin n'était pas dangereux.
<`[o2-mde-o1]`>

nous dirons que l'être dénoté par *celui que Pierre a tué ce matin*, à savoir un lion particulier, est `membre de` l'être dénoté par *Le lion*, à savoir la classe des lions. Le fait que l'expression *Le lion* dénote la classe des lions relève de ce que nous avons appelé « référence générique » (cf. section 1.4.4, page 37). Rappel : les cas de référence générique sont marqués dans la notation par l'ajout du symbole `(G)` à la suite de l'index identifiant le référent.

2.3.2 Définition

Nous avons vu (section 1.4.4) que nous distinguons dans l'univers de dénotation des êtres qui sont des êtres singuliers et des êtres qui sont des ensembles d'au moins deux êtres singuliers. Par exemple dans l'exemple (52) ci-dessus, nous dirons que l'expression *celui qu'elle préférait* dénote un être singulier et que l'expression *deux chevaux* dénote un ensemble d'être singuliers. Les notions d'« être singulier » et d'« ensemble d'êtres singuliers » dans notre langage de description des reprises correspondent respectivement à un atome et à un ensemble *d'au moins deux éléments* en théorie des ensembles. On remarque qu'il n'y a *a priori* rien qui soit équivalent aux singletons de la théorie des ensembles dans l'univers de dénotation tel que nous le concevons.

La proposition

$\mathtt{o}_i$ est membre de $\mathtt{o}_j$

fait sens dans le langage de description des reprises seulement si $\mathtt{o}_j$ est un ensemble d'êtres singuliers (on le répète, un ensemble d'êtres singuliers dans le langage de description des reprises correspond en théorie des ensembles à un ensemble ayant au moins deux éléments). Il n'y a aucune contrainte quant à la nature de $\mathtt{o}_i$ qui peut être soit un être singulier, soit un ensemble d'êtres singuliers.

La proposition du langage de description des reprises « $\mathtt{o}_i$ est membre de $\mathtt{o}_j$ » peut se traduire dans le langage de la théorie des ensembles. Elle est équivalente

à l'une ou l'autre des propositions suivantes, selon le type de l'être $\mathtt{o}_i$:

- si $\mathtt{o}_i$ est un être singulier, $\mathtt{o}_i$ *est membre de* $\mathtt{o}_j$ signifie dans le langage de la théorie des ensembles : « $\mathtt{o}_i \in \mathtt{o}_j$ »
- si $\mathtt{o}_i$ est un ensemble d'êtres singuliers, $\mathtt{o}_i$ *est membre de* $\mathtt{o}_j$ signifie dans le langage de la théorie des ensembles : « $\mathtt{o}_i \subset \mathtt{o}_j$ »

Pour pouvoir traduire une formule de la forme $\mathtt{o}_i$-`mde`-$\mathtt{o}_j$ (langage de description des reprises) dans le langage de la théorie des ensembles, on doit donc savoir si l'être $\mathtt{o}_i$ est un être singulier ou un ensemble d'êtres singuliers.

Cela étant, pour permettre au lecteur de bien cerner les différences entre notre relation `membre-de` et les deux relations d'appartenance et inclusion de la théorie des ensembles, nous proposons ici un exemple ou ces notions sont mises en correspondance. Soit un univers de dénotation complet contenant les êtres suivants :

- les êtres singuliers `o1`, `o2`, `o3`,
- et les ensembles d'êtres singuliers `o4` et `o5`.

Soient les relations suivantes entre ces êtres, notées suivant la notation indiquée plus haut :

$<$[o1-mde-o5]$>$
$<$[o2-mde-o5]$>$
$<$[o3-mde-o5]$>$
$<$[o4-mde-o5]$>$
$<$[o1-mde-o4]$>$
$<$[o2-mde-o4]$>$

Sachant pour chacun de ces êtres s'il est un être singulier ou un ensemble d'êtres singuliers, on peut inférer qu'en langage de la théorie des ensembles les extensions de `o4` et `o5` sont les suivantes :

$\mathtt{o4} = \{\mathtt{o1}, \mathtt{o2}\}$
$\mathtt{o5} = \{\mathtt{o1}, \mathtt{o2}, \mathtt{o3}\}$

Comme `o4` est un ensemble, la proposition du langage de description des reprises `o4` *est membre de* `o5` se traduit dans le langage de la théorie des ensembles par :

$\mathtt{o4} \subset \mathtt{o5}$

et non pas par :

$\mathtt{o4} \in \mathtt{o5}$

Cette dernière expression du langage de la théorie des ensembles n'est pas traduisible dans notre langage de description des reprises parce que `o4` est un ensemble. En général, une expression de la théorie des ensembles de la forme $\mathrm{A} \in \mathrm{B}$, n'est pas

traduisible dans notre langage de description des reprises si A est un ensemble.

Les cinq autres des expressions ci-dessus se traduisent dans le langage de la théorie des ensembles avec le symbole d'appartenance :

o1 $\in$ o5
o2 $\in$ o5
o3 $\in$ o5
o1 $\in$ o4
o2 $\in$ o4

Une fois traduite dans la théorie des ensembles, la relation `membre-de` ne donne lieu qu'à des ensembles dont les éléments sont des atomes.

2.3.3 Transitivité de la relation membre-de

La relation `membre-de` est transitive. Si les propositions suivantes sont vraies :

$<$[o_i-mde-o_j]$>$
$<$[o_j-mde-o_k]$>$

alors la proposition suivante est aussi vraie :

$<$[o_i-mde-o_k]$>$

On a en effet la traduction suivante dans le langage de la théorie des ensembles.

Si o_i-mde-o_j est vrai alors o_j est un ensemble d'êtres singuliers. Si o_j est un ensemble d'êtres singuliers, alors o_j-mde-o_k se traduit en langage de la théorie des ensembles par :

$o_j \subset o_k$

L'être o_i est soit un être singulier, soit un ensemble d'êtres singuliers. Dans le premier cas o_i-mde-o_j se traduit par :

$o_i \in o_j$

dans le second cas, par :

$o_i \subset o_j$

Le raisonnement ci-dessus peut donc être traduit de deux manières selon le type de o_i et donne lieu à deux raisonnements valides en théorie des ensembles :

- si $o_i \in o_j$ et $o_j \subset o_k$, alors $o_i \in o_k$
- si $o_i \subset o_j$ et $o_j \subset o_k$, alors $o_i \subset o_k$

Les deux cas ci-dessus étant les seuls possibles, la relation membre-de est transitive.

2.3.4 Relation « membre-de » et identité de description

Le lecteur attentif ou déjà versé dans ces questions aura remarqué que certains des exemples que nous avons utilisés pour présenter la relation `membre-de` ressemblaient fortement à certains des exemples où nous parlions d'identité de description. C'est que la relation `membre-de` implique une identité de description.

Il y a deux manières de définir un ensemble : en spécifiant chacun de ses éléments, par exemple :

$$E = \{1, 2, 3, 4, 5\}$$

ou en donnant une description de ses éléments, par exemple :

$$E = \{x \mid x \text{ est un entier positif inférieur à } 6\}$$

Étant donné l'ensemble E, que nous avons défini de deux manières différentes, on peut dire que la description *est un entier positif inférieur à 6* s'applique à chacun de ses éléments : il est vrai que 1 est un entier positif inférieur à 6, que 2 est également un entier positif inférieur à 6, etc.

De la même manière, si, dans notre langage de description des reprises, on dit d'un être $\mathtt{o}_i$ qu'il est `membre` d'un être $\mathtt{o}_j$, alors il doit y avoir une description commune à $\mathtt{o}_i$ et $\mathtt{o}_j$. Reprenons les exemples donnés au début de cette section.

(50) <Les manifestants `[o1]`> se sont dispersés, mais <la plupart `[o2]`> sont restées en ville.
<`[o2-mde-o1]`>

(51) Pierre avait <deux chevaux `[o1]`>. Marie a pris <le blanc `[o2]`>.
<`[o2-mde-o1]`>

(52) Pierre avait <deux chevaux `[o1]`>. Marie a pris <celui qu'elle préférait `[o2]`>.
<`[o2-mde-o1]`>

(53) Pierre et Marie ont <trois enfants `[o1]`> ; Jacques <en `[o1]`> connaît <deux `[o2]`>.
<`[o2-mde-o1]`>

Dans (50), l'être dénoté par *la plupart* est un ensemble de manifestants. Dans (51) et (52), les êtres dénotés respectivement par *le blanc* et *celui qu'elle préférait* sont tous deux des chevaux. Dans (53), l'être dénoté par *deux* est un ensemble d'enfants. Dans ces phrases, les descriptions des éléments de l'ensemble dénoté qui sont faites au niveau des expressions sources s'appliquent aux référents des expressions anaphoriques.

Nous avons tenté d'expliciter les différentes configurations permettant de caractériser la description qui est reprise au niveau d'une expression interprétée au

moyen de la relation `membre-de`. Ces configurations sont au nombre de trois. Elles diffèrent par le fait que l'expression qui dénote le sur-ensemble est :

1. un nom collectif au singulier ;
2. un nom non collectif au pluriel ;
3. un nom collectif au pluriel.

Un « nom collectif » est un nom qui au singulier est susceptible de dénoter un ensemble d'êtres homogènes quant à leur type. Cette classe de noms regroupe les noms tels que *troupeau, bande, groupe*, caractérisables par le fait qu'ils peuvent prendre comme complément un syntagme prépositionnel pluriel introduit par *de* et sans déterminant, syntagme tel que son noyau donne en fait une description des éléments de l'ensemble en question : *un troupeau de vaches, une bande de jeunes, un groupe de musiciens*.

Nous décrivons plus précisément ces trois configurations ci-dessous.

Soit $\mathtt{o}_j$ un ensemble et $\mathtt{o}_i$ un membre de cet ensemble. L'ensemble $\mathtt{o}_j$ est dénoté par un syntagme nominal sans coordination.

1. Si $\mathtt{o}_j$ est dénoté par un syntagme nominal dont le noyau est un nom collectif au singulier, alors la description qui est fait des éléments qui composent l'ensemble $\mathtt{o}_j$ s'applique aux éléments de $\mathtt{o}_i$. Cette description peut être explicitée par un syntagme prépositionnel rattaché au syntagme qui dénote $\mathtt{o}_j$ (exemples (55) et (56)) ou être une description par défaut associée au nom collectif (exemples (57)).

(55) <Une foule `[o1]`> de chiens a envahi la ville. <Quelques labradors `[o2]`> ont mordu le maire.
<`[o2-mde-o1]`>

La description *chien* s'applique aux êtres de l'ensemble dénoté par *Quelques labradors*.

(56) <Le régiment `[o1]`> de parachutistes se dispersa. <Certains `[o2]`> restèrent en ville.
<`[o2-mde-o1]`>

La description *parachutiste* s'applique aux êtres de l'ensemble dénoté par *Certains*.

(57) <La foule `[o1]`> se dispersa. <Certains `[o2]`> restèrent en ville.
<`[o2-mde-o1]`>

Au nom collectif *foule* est associé par défaut la description *personne* ; celle-ci s'applique aux êtres de l'ensemble dénoté par *Certains*.

2. Si $\mathtt{o}_j$ est dénoté par un syntagme nominal dont le noyau est un nom non collectif au pluriel, alors la description qui est fait des éléments qui composent l'ensemble $\mathtt{o}_j$ s'applique aux éléments de $\mathtt{o}_i$.

(58) <Les manifestants [o1]> se dispersèrent, mais <les meneurs [o2]> restèrent en ville.
<[o2-mde-o1]>

La description *manifestant* s'applique aux êtres de l'ensemble dénoté par *les meneurs*.

3. Si o_j est dénoté par un syntagme nominal dont le noyau est un nom collectif au pluriel, et si aucune description des individus qui composent o_i n'est fournie, alors il y a potentiellement ambiguïté sur la description qui est reprise de la source.

(59) <Les régiments [o1]> de parachutistes se dispersèrent. <Certains [o2]> restèrent en ville, <d'autres [o3]> partirent vers le nord.
<[o2-mde-o1]>
<[o3-mde-o1]>

La description qui s'applique à l'être dénoté par *Certains* peut être *régiment* ou *parachutiste*, et de même pour le syntagme *d'autres*.

(60) <Les régiments [o1]> se dispersèrent. <Certains [o2]> restèrent en ville, <d'autres [o3]> partirent vers le nord.
<[o2-mde-o1]>
<[o3-mde-o1]>

La description dont hérite l'être dénoté par *Certains* peut être *régiment*, mais aussi une description associée par défaut aux individus qui composent les régiments, par exemple *soldat*.

2.3.5 Relation membre-de et identité de dénotation

Pour terminer cette présentation de la relation `membre-de`, il nous faut la mettre en rapport avec les cas de reprises avec identité de dénotation présentés page 45 (section 2.1.3). Nous reproduisons ici un exemple d'une telle reprise :

(8) La pression monte chaque jour d'un cran à Milan pour pousser à un mariage entre <la Comit [o1]> et <Banca di Roma [o2]>, qui donnerait naissance à la première banque italienne. Les informations s'accumulent, laissant penser que le processus visant à préparer la fusion est lancé. Ainsi, <les deux banques [o3]> ont déjà choisi leurs banques conseils (Merrill Lynch pour la Comit et Goldman Sachs pour Banca di Roma).
<[{o1,o2}-id-o3]>

Cet exemple pourrait aussi être analysé en utilisant la relation `membre-de` :

(8) La pression monte chaque jour d'un cran à Milan pour pousser à un mariage entre <la Comit [o1]> et <Banca di Roma [o2]>, qui donnerait naissance à la première banque italienne. Les informations s'accumulent, laissant penser que le processus visant à préparer la fusion est lancé. Ainsi, <les deux banques [o3]> ont déjà choisi leurs banques conseils (Merrill

Lynch pour la Comit et Goldman Sachs pour Banca di Roma).
<[o1-mde-o3]>
<[o2-mde-o3]>

Les emplois potentiellement concurrents des notations $\texttt{o}_i\texttt{-id-o}_j$ et $\texttt{o}_i\texttt{-mde-o}_j$ sont réglementés par le principe suivant :

– la notation $\{\texttt{o}_j,\ldots,\texttt{o}_n\}\texttt{-id-o}_i$ ne peut être utilisée que si l'extension de l'ensemble o_i est entièrement spécifiée par $\{\texttt{o}_j,\ldots,\texttt{o}_n\}$ et si la première mention de l'être $\texttt{o}_i$ suit les expressions $\texttt{o}_j$,..., $\texttt{o}_n$.

La notation à utiliser pour l'exemple (8) ci-dessus est donc la première des deux notations envisagées.

En revanche, dans l'exemple suivant, on utilisera la relation `membre-de`, dans la mesure où la première mention de l'ensemble `o1` précède la première mention de ses éléments.

(61) Le secteur bancaire sud-coréen a subi hier un nouveau choc, avec l'annonce par l'agence américaine Moody's de sa décision de dégrader les notations des <trois principales banques [o1]> du pays contrôlées par l'État. Motif invoqué : le plan de soutien du gouvernement sud-coréen à <ces banques [o1]>, destiné à garantir <leur [o1]> dette internationale, n'est pas suffisamment explicite. Dans le détail, Moody's a revu de « Ba1 » à « Ba2 » les notes de la dette à long terme de <la Korea Development Bank [o2]> (KDB), principale banque du pays, et de <l'Export-Import Bank of Korea [o3]>.
<[o2-mde-o1]>
<[o3-mde-o1]>
L'agence a été encore plus sévère pour <l'Industrial Bank of Korea [o4]> (IBK) spécialisée dans le financement des PME-PMI, <qui [o4]> a vu <sa [o4]> notation abaissée de deux crans, de « Ba1 » à « Ba3 ».
<[o4-mde-o1]>
<Ces trois banques [o1]> sont utilisées par le gouvernement pour tenir à bout de bras un grand nombre d'entreprises en difficulté, et ce « en mésestimant <leur [o1]> solidité financière intrinsèque », explique encore Moody's qui relève, à titre d'illustration, que <la KDB [o2]> porte un endettement à moyen et long terme de l'ordre de 17 milliards de dollars.

2.4 Relation « distingué-de »

Notre quatrième type de reprise met en jeu l'emploi de l'adjectif *autre*. Ce type de reprise met en jeu une identité de description, mais il est intéressant de le distinguer de la simple identité de description décrite dans la section 2.2 dans la mesure où l'adjectif *autre* explicite la négation de l'identité de dénotation.

Comme pour la relation `membre-de`, qui, outre d'une identité de description, met en jeu une identité de dénotation partielle, nous avons dans les reprises avec *autre* un type de reprise plus spécifique que la seule identité de description.

L'emploi de *autre* implique une différenciation par rapport à quelque chose. Pour parler d'un y différent de x, il faut disposer de x. Il s'ensuit qu'un syntagme nominal avec *autre* dénotant un être $\mathtt{o}_i$ sera très souvent analysable comme une reprise par rapport à un autre syntagme nominal dénotant un être $\mathtt{o}_j$, en vertu précisément de la présence de *autre*.

Sur un ensemble de 44 occurrences de *autre* dans un recueil d'articles de La Tribune, nous avons relevé :

- 6 tournures plus ou moins idiomatiques (*d'autre part, en d'autres termes, de part et d'autre de...*),
- 2 emplois de *autre* dans un contexte négatif (*rien d'autre, pas d'autre choix que de...*),
- 35 occurrences de *autre* à l'intérieur d'un syntagme nominal dénotant, telles qu'il était possible de localiser dans le texte une expression (ou un ensemble d'expressions) qui justifie l'emploi de *autre*,
- et un dernier cas où *autre*, dans un syntagme nominal dénotant, renvoyait plutôt à la situation d'énonciation (nous reviendrons sur ce cas ci-dessous).

Étant donné un syntagme nominal avec *autre*, dénotant un être $\mathtt{o}_i$, nous dirons que ce syntagme est interprété par reprise s'il existe dans le texte qui précède une expression dénotant $\mathtt{o}_j$ et si $\mathtt{o}_i$ est décrit comme *autre* par rapport à $\mathtt{o}_j$. Nous dirons de $\mathtt{o}_i$ qu'il est « distingué » de $\mathtt{o}_j$ et nous appellerons la relation qui lie ces deux expressions la relation `distingué-de`, notée $\mathtt{o}_i$`-dde-`$\mathtt{o}_j$.

2.4.1 Exemples

Dans la phrase suivante, nous disons que l'être dénoté par *autre* est distingué de l'être dénoté par *un État*.

(62) Pour la CJCE, l'objet d'une convention n'est pas de garantir au contribuable que l'imposition due dans <un État `[o1]`> ne soit pas supérieure à celle qu'il doit payer dans <l'autre `[o2]`>.
<`[o2-dde-o1]`>

Dans la phrase suivante, l'être dénoté par *les dix-neuf autres commissaires européens* est distingué de l'être dénoté par *Karel Van Miert*.

(63) Le compromis laborieusement négocié depuis le 1er mai entre les proches collaborateurs de Dominique Strauss-Kahn et de <Karel Van Miert `[o1]`> a finalement été entériné, hier, par les cabinets des <dix-neuf autres commissaires européens `[o2]`>.
<`[o2-dde-o1]`>

Dans la phrase suivante, l'être dénoté par *d'autres établissements* est distingué de l'être dénoté par *Le Crédit Agricole*.

(64) <Le Crédit Agricole `[o1]`> va perdre le monopole des dépôts des notaires. Le ministère des Finances pourrait ouvrir la collecte de ces fonds en milieu rural à <d'autres établissements `[o2]`>.
<`[o2-dde-o1]`>

2.4.2 Définition

Nous avons indiqué ci-dessus dans quelles conditions et comment nous décrivons des reprises avec *autre*, reprise mettant en jeu une relation que nous avons appelée `distingué-de`. Nous précisons ici ce qu'est cette relation en la mettant en parallèle avec la théorie des ensembles, comme nous l'avons fait pour la relation `membre-de`.

Certains cas de reprise avec *autre* peuvent être envisagés en termes de complémentarité entre deux ensembles relativement à un ensemble de référence. En théorie des ensembles, l'opération *complément* se définit par rapport à un *ensemble universel* (ou référentiel) U. Le complément d'un ensemble A, que nous désignerons $\mathcal{C}(A)$, est l'ensemble de tous les éléments de U qui n'appartiennent pas à A :

$$\mathcal{C}(A) = \{x \mid x \in U \text{ et } x \notin A\}$$

Dans le discours,

(65) La foule se dispersa. Quelques personnes restèrent en ville, les autres rentrèrent chez elles.

on peut interpréter l'ensemble dénoté par *les autres* comme le complémentaire ($\mathcal{C}(A)$) de l'ensemble dénoté par *Quelques personnes* (A) sur l'ensemble dénoté par *la foule* (U).

Toutes les reprises avec *autre* ne peuvent cependant s'interpréter en terme de complémentarité. Dans l'exemple suivant, les êtres dénotés par *l'un d'entre eux* et *un autre* sont deux êtres singuliers distincts tels que chacun est `membre-de` l'ensemble dénoté par *ces candidats potentiels*, mais ils ne sont pas complémentaires sur cet ensemble.

(66) Aucun de ces candidats potentiels n'aborde le dossier avec un enthousiasme débordant. « Nous y allons parce que nous regardons systématiquement les opportunités et que nous commettrions une faute en ne regardant pas la SMC », explique l'un d'entre eux. « Nous y allons comme tout le monde par curiosité, pour ne pas être absent », précise un autre.

Il n'y a pas ici complémentarité, mais la notion peut quand même nous être utile. Supposons que le syntagme *ces candidats potentiels* fasse référence à l'être `o1`, un ensemble, et que `o1` soit composé des êtres singuliers `o2`, `o3`, `o4` et `o5`. Le syntagme *l'un d'entre eux* fait alors référence soit à `o2`, soit à `o3`, soit à `o4`, soit à `o5`. La

fonction principale de l'adjectif *autre* nous paraît être de distinguer l'être dénoté par *un autre* par rapport à celui qui est dénoté par *l'un d'entre eux*. En d'autres termes, si *l'un d'entre eux* dénote, par exemple, `o2`, alors *un autre* dénote soit `o3`, soit `o4`, soit `o5`, or ces trois êtres sont précisément ceux qui composent $\mathcal{C}$(`o2`), l'ensemble complémentaire de `o2` sur `o1`, c'est-à-dire que le référent de *un autre* doit être `membre-de` $\mathcal{C}$(`o2`).

On se donne donc une nouvelle notion, la notion de « distinction », qui se définit par rapport à la complémentarité. Étant donné :

- un ensemble référentiel U,
- un ensemble A,
- et $\mathcal{C}(A)$, le complément de A sur U,

on dit d'un être $\mathtt{o}_i$ qu'il est `distingué-de` $\mathtt{o}_j$ s'il est `membre-de` $\mathcal{C}(\mathtt{o}_j)$. Par définition, $\mathcal{C}(\mathtt{o}_j)$ est `distingué-de` $\mathtt{o}_j$. La complémentarité d'un ensemble par rapport à un autre est donc un cas particulier de distinction. Dans la mesure où il semble qu'on puisse déterminer s'il y a complémentarité ou non selon que le syntagme avec *autre* contient ou ne contient pas un article défini [11], on n'envisage pas d'adopter une notation particulière pour ce cas.

Le plus souvent, dans les exemples que nous avons observés, il n'est pas fait explicitement référence à l'ensemble de référence en fonction duquel la complémentarité, et donc la distinction, sont exprimées. En revanche, comme nous l'avons signalé, il y a presque toujours référence à un être dont le référent de l'expression avec *autre* est distingué. C'est la relation du référent de l'expression avec *autre* avec cet être qui constitue l'élément constant de la reprise avec *autre*, et c'est cette relation qui est principalement notée. La relation `membre-de` est opérationnelle pour les cas où l'ensemble de référence est exprimé.

La notation pour les exemples (65) et (66) ci-dessus sera donc la suivante :

(65) <La foule `[o1]`> se dispersa. <Quelques personnes `[o2]`> restèrent en ville, <les autres `[o3]`> rentrèrent chez elles.
<`[o2-mde-o1]`>
<`[o3-mde-o1]`>
<`[o3-dde-o2]`>

(66) Aucun de <ces candidats potentiels `[o1]`> n'aborde le dossier avec un enthousiasme débordant. « Nous y allons parce que nous regardons systématiquement les opportunités et que nous commettrions une faute en ne regardant pas la SMC », explique <l'un `[o2]`> d'entre <eux `[o1]`>. « Nous y allons comme tout le monde par curiosité, pour ne pas être absent », précise <un autre `[o3]`>.
<`[o2-mde-o1]`>
<`[o3-mde-o1]`>
<`[o3-dde-o2]`>

[11] À ce sujet, voir, par exemple, [82, p. 204].

2.4.3 Absence de source explicite pour une reprise avec *autre*.

Nous avons évoqué un cas où aucune expression ne fait explicitement référence à un être par rapport auquel le référent d'une expression avec *autre* serait distingué. Dans le texte où apparaît la phrase suivante, l'expression *les autres pays* désigne un ensemble de pays autres que la France, sans qu'il soit explicitement fait référence à la France dans le texte.

(67) Soit les choses se font de façon correcte, comme dans les autres pays où l'on observe une rémunération modérée, généralement inférieure ou égale à 0,5 % et accompagnée d'une facturation modérée des chèques.

Dans un tel cas de reprise avec *autre* sans référence explicite à un être qui justifie l'emploi de *autre*, on note au tout début du texte une séquence de la forme :

<[*description*, o_n]>

Cette séquence est destinée à indiquer l'être en fonction duquel le référent décrit par *autre* est distingué. On indique entre crochets carrés une description du référent, au moyen d'une chaîne de caractères correspondant à un nom propre ou à un lemme, et un index de référent (o_n) qui sera utilisé pour noter la relation `distingué-de` au niveau de la reprise. Pour notre exemple, on aurait ainsi :

(67) <[France, o1]>
début du texte...
... Soit les choses se font de façon correcte, comme dans <les autres pays [o2]> où l'on observe une rémunération modérée, généralement inférieure ou égale à 0,5 % et accompagnée d'une facturation modérée des chèques.
<[o2-dde-o1]>

2.4.4 Relation « distingué-de » et identité de description

Comme pour la relation `membre-de`, il y a dans les reprises avec *autre* une identité de description entre les êtres dénotés par les deux expressions en relation. Nous avons tenté de déterminer quelle était la description commune aux deux référents dans différentes configurations.

Supposons deux êtres o_i et o_j, dénoté respectivement par une expression e_i et une expression e_j et tels que o_j est décrit comme *autre* par rapport à o_i. Nous distinguerons les cas de figure suivants :

1. *autre* est le noyau de e_j et il n'existe pas d'expression dans le texte qui dénote un être o_k dont o_i et o_j seraient tous deux membres ;
2. *autre* est le noyau de e_j et il existe une expression e_k qui dénote o_k dont o_i et o_j sont tous deux membres ;
3. *autre* n'est pas le noyau de e_j.

1. Si *autre* est le noyau de e_j et s'il n'existe pas d'expression dans le texte qui dénote à un être o_k dont o_i et o_j seraient tous deux membres, alors la description

nominale qui est faite de $\mathtt{o}_i$ dans e_i s'applique à $\mathtt{o}_j$, ou bien, si $\mathtt{o}_i$ est un ensemble dénoté par une coordination de syntagmes nominaux, la description commune aux deux êtres est implicite. Exemple du premier cas :

(68) <Une personne [$\mathtt{o}_i$]> a apporté des fleurs ; <une autre [$\mathtt{o}_j$]> a apporté des bonbons.
<[$\mathtt{o}_j$-dde-$\mathtt{o}_i$]>

Exemple pour le second cas :

(69) Seuls <Pierre et Juliette [$\mathtt{o}_i$]> sont venus. <Les autres [$\mathtt{o}_j$]> étaient en vacances.
<[$\mathtt{o}_j$-dde-$\mathtt{o}_i$]>

Les êtres qui composent l'ensemble $\mathtt{o}_j$ sont de même type que les êtres Juliette et Pierre, à savoir des personnes. Cette description commune est ici implicite.

2. Si *autre* est le noyau de e_j et s'il existe une expression e_k qui dénote $\mathtt{o}_k$ dont $\mathtt{o}_i$ et $\mathtt{o}_j$ sont tous deux membres, alors, en vertu de ce que nous avons indiqué pour la relation `membre-de` (voir page 61), la description, dans e_k, des êtres qui composent $\mathtt{o}_k$ s'applique aux êtres $\mathtt{o}_i$ et $\mathtt{o}_j$.

La description nominale qui est faite de $\mathtt{o}_i$ dans e_i peut éventuellement s'appliquer aussi à $\mathtt{o}_j$ si e_i contient un complément, comme dans l'exemple (70), où l'adjectif *autre* nie la description *les plus enragés* (c'est-à-dire que les autres sont des meneurs, mais pas les plus enragés), ou si la description présente dans e_i est la même que celle qui est présente, explicitement ou implicitement, dans e_k, comme dans l'exemple (71). L'exemple (72) illustre un cas où la description nominale de e_i ne s'applique pas à $\mathtt{o}_j$.

(70) <Les manifestants [$\mathtt{o}_k$]> se dispersèrent. <Les meneurs [$\mathtt{o}_i$]> les plus enragés restèrent en ville, <les autres [$\mathtt{o}_j$]> rentrèrent chez eux.
<[$\mathtt{o}_i$-mde-$\mathtt{o}_k$]>
<[$\mathtt{o}_j$-mde-$\mathtt{o}_k$]>
<[$\mathtt{o}_j$-dde-$\mathtt{o}_i$]>

(71) <La foule [$\mathtt{o}_k$]> se dispersa. <Quelques personnes [$\mathtt{o}_i$]> restèrent en ville, <les autres [$\mathtt{o}_j$]> rentrèrent chez elles.
<[$\mathtt{o}_i$-mde-$\mathtt{o}_k$]>
<[$\mathtt{o}_j$-mde-$\mathtt{o}_k$]>
<[$\mathtt{o}_j$-dde-$\mathtt{o}_i$]>

(72) <Les manifestants [$\mathtt{o}_k$]> se dispersèrent. <Les meneurs [$\mathtt{o}_i$]> restèrent en ville, <les autres [$\mathtt{o}_j$]> rentrèrent chez eux.
<[$\mathtt{o}_i$-mde-$\mathtt{o}_k$]>
<[$\mathtt{o}_j$-mde-$\mathtt{o}_k$]>
<[$\mathtt{o}_j$-dde-$\mathtt{o}_i$]>

3. Si *autre* n'est pas le noyau de e_j, alors la description nominale qui est faite de $\mathtt{o}_j$ dans e_j s'applique à l'être $\mathtt{o}_i$. Si les deux situations précédentes mettaient

en jeu une ellipse du noyau nominal du syntagme qui nécessitait une reprise de description, dans la situation présente, c'est l'expression source qui hérite d'une description présente au niveau de la reprise. Les exemples (63) et (64), page 65, illustrent cette situation ; en voici un autre :

(73) Dans cette mesure, l'article 6.1 de la Convention européenne, qui énonce que <le citoyen peut saisir pour toute décision prise à son encontre « un tribunal offrant les garanties de ce texte » [o_i]>, est incompatible avec le texte du CGI. <Un autre principe [o_j]> énoncé par la Convention européenne et employé par la cour suprême en 1997 relève de « l'égalité des armes ».
<[o_j-dde-o_i]>

Le fait que tout citoyen puisse saisir pour toute décision prise à son encontre « un tribunal offrant les garanties de ce texte » (être o_i) est un principe, comme l'est l'être o_j..

Pour terminer, on remarquera que la communauté de description entre les êtres qui sont distingués avec *autre* semble être à l'origine de l'effet produit par des tournures telles que :

(74) les Pompidou, Giscard et autres Chirac

tournures dans lesquelles on cherche à signifier que tous les êtres dénotés sont de même type.

2.5 Référence à un discours

Le cinquième et dernier type de reprise que nous distinguons est plus particulier que les précédents dans la mesure où il se limite aux reprises par un pronom clitique objet d'un verbe de discours. Considérons le discours :

(75) Jacques est parti au cinéma. Marie vient de me le dire.

On a là un lien de reprise qui n'est ni l'identité de description, ni l'identité de dénotation. Nous considérons que l'expression *Jacques est parti au cinéma* dénote (et décrit) un être de l'univers de dénotation `o1`. Le clitique *le* ne peut être vu comme dénotant l'être `o1` : ce que Marie a dit ce n'est pas l'être `o1` à proprement parler, mais une description de cet être. L'objet d'un verbe de parole comme *dire* est, par définition, une description. Cependant le clitique *le* ne reprend pas à proprement parler la description *Jacques est parti au cinéma*. En effet, Marie a pu faire une description de `o1` différente de la description *Jacques est parti au cinéma* ; elle a pu dire par exemple *Mon frère est allé voir le dernier film de Woody Allen*. Nous avons donc la situation suivante :

- on suppose dans l'univers de dénotation un être `o1`,
- il y a une description de `o1` : *Jacques est parti au cinéma*,

– il y a une référence à une description de `o1`, qui n'est pas nécessairement la même que la précédente (c'est-à-dire la description qu'a dite Marie).

Nous dirons que chacune des deux descriptions en jeu dans ce type de situation est une « paraphrase » de l'autre. Deux descriptions sont des paraphrases si elles décrivent le même être tout en étant différentes.

La notation adoptée est la suivante. Le lien entre les deux descriptions paraphrases étant l'être qu'elles dénotent, on utilise l'index de référent. On note au niveau de la source un index de référent :

(75) <Jacques est parti au cinéma `[o1]`>. Marie vient de me le dire.

Comme nous l'avons dit, le clitique *le* ne reprend ni le référent proprement dit, ni la description du référent, mais désigne une description qui est une paraphrase de la description de `o1` qui apparaît dans le texte. On utilise au niveau de la reprise la notation $\texttt{o}_i\texttt{(P)}$, qui se lit « paraphrase dénotant $\texttt{o}_i$ ».

(75) <Jacques est parti au cinéma `[o1]`>. Marie vient de me <le `[o1(P)]`> dire.

Avec ce type de reprise, sur lequel nous reviendrons au chapitre 4, s'achève la présentation de notre typologie des reprises. Nous évoquons dans les deux sections suivantes quelques phénomènes d'anaphore associative et de deixis, que nous avons écartés de notre définition des reprises au chapitre 1 (voir en particulier la section 1.2).

2.6 Au-delà des reprises : anaphore associative et deixis

Nous avons, dans le premier chapitre, restreint la notion de reprise aux liens caractérisés en ayant recours à une relation d'identité parce que nous pensons, au vu des résultats de l'expérience décrite au chapitre 4, que ces liens pourront être observés de manière inter-subjective et que cette inter-subjectivité, même si elle reste à démontrer, autorise un élargissement de la notion d'anaphore à un ensemble de liens qui ne sont pas caractérisés par rapport à la forme des expressions.

Pour l'expérience qui sera décrite au chapitre 4, nous avons demandé à un groupe d'observateurs de noter, outre les liens de reprise, l'ensemble des relations qui peuvent être observées dans les cas d'anaphore associative, ainsi que quelques phénomènes relevant de la deixis, en l'occurrence l'interprétation des expressions qui font référence à une date. Nous donnons ici la description de ces phénomènes, telle que proposée pour l'expérience du chapitre 4. Cela nous permettra, d'une part, de mettre en perspective la notion de reprise telle que définie dans la section 1.2 avec des phénomènes qui lui sont très proches, d'autre part, de donner les éléments qui permettront de comprendre l'expérience décrite par la suite.

2.6.1 Relations référentielles

Au-delà des reprises, on peut observer que des expressions dénotent des êtres qui entretiennent entre eux une relation sans qu'il y ait un lien d'identité de description. C'est le cas dans les phénomènes qu'on caractérise habituellement comme relevant de l'« anaphore associative » : un syntagme nominal défini dénote un être dont l'existence est déduite de l'existence d'un autre être désigné auparavant dans le discours. Nous caractérisons ce type de situation comme mettant en jeu une « relation référentielle », c'est-à-dire une relation entre référents.

Nous n'avons pas développé la caractérisation des différentes relations qui peuvent être observées entre deux êtres de l'univers de dénotation. Pour le test d'opérationalité qui sera décrit au chapitre 4, nous avons simplement caractérisé une de ces relations, la relation `partie-de`, et utilisé une notation à la sémantique large pour rendre compte des relations qui pouvaient être observées entre deux êtres sans que celle-ci soit une des relations précédemment définies.

Relation « partie-de »

Une expression dénotante peut être interprétée de telle manière qu'elle dénote un être qui est une partie d'un être de l'univers de dénotation identifié préalablement. On se donne une relation `partie-de`, notée $\mathtt{o}_i$`-pde-`$\mathtt{o}_j$, pour mettre en relation ces deux référents.

Notre relation `partie-de` a globalement la même sémantique que l'expression française « être une partie de » [12]. On utilisera donc une formule de la forme $\mathtt{o}_i$`-pde-`$\mathtt{o}_j$ si $\mathtt{o}_i$ dénote une pièce, un organe entrant dans la composition de $\mathtt{o}_j$ (par exemple, la roue d'un vélo, un membre ou un organe d'un être humain). L'être $\mathtt{o}_i$ sera dit aussi `partie-de` $\mathtt{o}_j$ s'il est un événement ou un processus qui se déroule à l'intérieur d'un processus plus large (par exemple, le vote lors d'une élection).

Dans le texte suivant, l'expression *la cathédrale* est interprétée comme désignant un être qui est une partie de l'être dénoté par *Clermont-Fd*.

(76) <Jean `[o1]`> est allé à <Clermont-Fd `[o2]`>. <Là `[o2]`>, <il `[o1]`> a visité <la cathédrale `[o3]`>.
<`[o3-pde-o2]`>

Dans la phrase suivante, l'être dénoté par *le câble d'alimentation* est une partie de l'être dénoté par *l'imprimante*.

(77) Ne placez pas <l'imprimante `[o1]`> à un endroit où des personnes pourraient marcher sur <le câble `[o2]`> d'alimentation.
<`[o2-pde-o1]`>

[12] Il importe de ne pas se laisser abuser par les différents emplois du nom *partie* en français : le mot pourra être utilisé pour mettre en relation un être qui est `membre` d'un autre être, comme dans la phrase *Jacques Chirac fait partie des présidents de la Ve République.*

Dans la phrase suivante, l'être dénoté par *la branche Chase Global Investor Services* est `partie-de` l'être dénoté par *la Chase*.

(78) Les activités et le personnel de Morgan Stanley seront intégrés à <la branche [o1]> Chase Global Investor Services, l'activité mondiale de <la Chase [o2]> dans la conservation de titres et la compensation.
<[o2-pde-o1]>

Comme nous l'avons signalé, la relation `partie-de` ne se limite pas aux seuls êtres de l'univers de dénotation qui peuvent être conçus comme des entités. Un « événement » peut être une partie d'un autre événement. C'est le cas de l'événement dénoté par *les négociations* dans la phrase suivante ; il est `partie-de` l'événement dénoté par *la transaction* :

(79) Lors des <négociations [o1]>, la presse avait rapporté que <la transaction [o2]> <s' [o2]> élevait autour de 600 millions de dollars.
<[o1-pde-o2]>

La relation `partie-de` présente l'intérêt d'avoir une sémantique assez proche de `membre-de`, proximité qui pourra donner lieu à d'éventuelles interrogations. La comparaison de ces deux relations nous permettra d'insister sur la signification de la relation `membre-de`, et plus généralement sur la notion de reprise : la différence entre les deux relations réside dans le fait que, contrairement à la relation `membre-de`, la relation `partie-de` n'implique pas d'identité de description.

Choisir des relations $\mathtt{o}_i$-mde-$\mathtt{o}_j$ plutôt que $\mathtt{o}_i$-pde-$\mathtt{o}_j$ dans les exemples ci-dessus aurait impliqué une vision de l'être $\mathtt{o}_j$ comme un ensemble d'être singuliers caractérisables par une description d_i, qui s'applique à l'être $\mathtt{o}_i$. Pour l'exemple (76), cela aurait impliqué que nous voyions Clermont-Fd comme un ensemble de bâtiments et la cathédrale de Clermont-Fd comme l'un de ces bâtiments — *bâtiment* étant la description d_i qui s'applique à l'être dénoté par la cathédrale. Pour l'exemple (77), cela aurait impliqué que nous voyions l'imprimante comme un ensemble de parties dont l'une serait le câble d'alimentation — *partie* étant notre description d_i. En d'autres termes, cela aurait consisté à voir les expressions *Clermont-Fd* et *imprimante* comme des sortes de noms collectifs. Tout être de l'univers de dénotation peut *a priori* être vu comme un ensemble d'éléments auxquels ont peut appliquer une même description. Le choix de la relation `membre-de` ou `partie-de` dépendra du niveau de généralité qu'on veut donner à cette description.

Exemples d'autres relations référentielles

Au-delà de la relation `partie-de`, nous présentons quelques exemples de relation entre êtres de l'univers de dénotation. Ces relations sont notées par une suite $\mathtt{o}_i$-rel-$\mathtt{o}_j$, qui signifie que les êtres $\mathtt{o}_i$ et $\mathtt{o}_j$ entretiennent entre eux une relation qui n'est ni une relation `partie-de`, ni une relation mettant en jeu une identité.

Dans le texte suivant, les « possibles conséquences » évoquées par Louis Viannet sont des conséquences de l'accord trouvé entre Paris et Bruxelles pour le Crédit Lyonnais. Par ailleurs, les éléments en possession de la CGT sont des éléments concernant le Crédit Lyonnais, l'être dénoté par *les éléments en notre possession* est donc en relation avec l'être dénoté par *le Crédit Lyonnais*. Enfin, les actifs dont parle Louis Viannet sont des actifs du Crédit Lyonnais [13].

(80) À peine confirmé, <l'accord `[o1]`> de principe trouvé entre Paris et Bruxelles pour <le Crédit Lyonnais `[o2]`> suscite déjà des remous. Le secrétaire général de la CGT, Louis Viannet, a estimé vendredi que le « diktat » du commissaire européen à la concurrence l'avait « emporté ». Dans une lettre au Premier ministre Lionel Jospin, le leader syndical met en garde contre <de possibles conséquences `[o3]`> « catastrophiques » pour l'emploi.
<`[o3-rel-o1]`>
« <Les éléments `[o4]`> en notre possession font en effet apparaître un total de cession d'<actifs `[o5]`> en Europe, dans le monde et en France d'environ 800 milliards », poursuit-il.
<`[o4-rel-o2]`>
<`[o5-rel-o2]`>

Dans l'exemple suivant, il est question de la privatisation du GAN, de la privatisation de la SMC et de la privatisation du Crédit Foncier. Il est fait référence à l'ensemble d'êtres singuliers contenant ces trois événements par le syntagme *les autres dossiers financiers aujourd'hui sur la rampe de lancement* et par le syntagme *ces trois ventes*. Chacune des références aux trois êtres qui sont vendus est considérée comme étant en relation avec cet ensemble.

(81) Fort de son succès dans la privatisation du CIC, le gouvernement garde l'ambition de boucler rapidement <les autres dossiers `[o1]`> financiers aujourd'hui sur la rampe de lancement. Qu'il s'agisse du <GAN `[o2]`>, du <Crédit Foncier de France `[o3]`> et de <la Société Marseillaise de Crédit `[o4]`>, Bercy garde bon espoir de boucler <ces trois ventes `[o1]`>, de gré à gré d'ici au début juillet.
<`[{o2,o3,o4}-rel-o1]`>

2.6.2 Référence à une date

La très grande majorité des liens de reprise, ainsi que les relations référentielles dont nous avons donné des exemples dans la section précédente, mettent en jeu des expressions dénotantes. L'identification des liens de reprise et des relations

[13] N.B. Toutes les interprétations par reprise ne sont pas notées dans les extraits suivants.

référentielles est intéressante dans la mesure où elle permet dans une certaine mesure de spécifier l'univers de dénotation associé à un texte (c'est-à-dire identifier les différents êtres désignés et les description qui en sont faites).

Un problème non couvert par l'idée d'une relation entre expressions est celui des expressions « déictiques ». Les expressions déictiques sont des expressions en général sous-spécifiques (p. ex. *aujourd'hui* peut être utilisé pour désigner n'importe quel jour) et qui sont interprétées non pas en fonction du discours mais en fonction de la situation d'énonciation du discours. Les expressions déictiques sont un problème parce que, comme leur référent n'est pas désigné par ailleurs par une expression du texte, on ne dispose pas d'un moyen direct de les représenter.

À titre de première piste vers ce que pourrait être un traitement de la deixis, nous proposons dans cette section un moyen de rendre compte de l'interprétation d'un sous-ensemble des expressions déictiques, en l'occurrence celles qui font référence à une date. La raison du choix de ces expressions est qu'on dispose d'un langage précis (les noms de jours, de mois, les numéros des années) pour décrire avec une spécificité maximale les êtres qu'elles désignent.

On peut faire l'analogie entre les dates et les noms propres. De la même manière qu'un nom propre, les expressions *1999*, *février 1999*, *le 25 février 1999* désignent de manière univoque et indépendante du contexte respectivement une année, un mois et un jour déterminés. On peut considérer de telles expressions comme des descriptions complètes de dates ou comme des « désignateurs rigides » de dates.

On peut cependant désigner de manière univoque une date (une année, un mois, un jour, mais aussi un trimestre, une semaine, etc) avec une expression qui s'interprète en fonction du contexte. Si en février 1999, je parle du « mois dernier », je désigne le mois de janvier 1999. Si le 25 février 1999, je parle de « demain », je désigne le 26 février 1999.

L'emploi de telles expressions sous-spécifiques pour désigner des dates déterminées est fréquent dans les articles de journaux. Dans le texte suivant, les expressions *vendredi* et *février* désignent respectivement un jour et un mois déterminés, mais ces référents ne peuvent être identifiés qu'avec des informations sur la situation d'énonciation, en l'occurrence la date de publication de l'article.

(82) La Commission européenne a approuvé *vendredi* l'acquisition des AGF par Allianz. Le feu bruxellois lève ainsi le dernier obstacle formel à la contre-offre publique d'achat (OPA) lancée en *février* sur l'assureur français.

L'article dont sont extraites ces lignes ayant été publié le lundi 11 mai 1998, l'expression *vendredi* désigne le vendredi 8 mai 1998 et l'expression *février* le mois de février 1998.

Pour désigner une date, l'énonciateur est donc susceptible d'employer une description complète ou une description incomplète. Le tableau 2.1 indique ce

Type de date	*Description complète*	*Exemple*
année	numéro de l'année	*1999, l'an 2000*
mois	nom du mois + année	*février 1999*
jour	numéro + mois + année	*le 20 mai 1998*

TAB. 2.1 – Modèles de description complète pour les dates

qu'on pense être la description complète pour les trois principaux types de dates.

Pour le test d'opérationalité décrit au chapitre 4, nous avions demandé aux observateurs de suivre la démarche suivante, lorsqu'ils rencontraient une référence à une date effectuée avec une description incomplète :

- Si une description complète de la date en question a déjà été fournie auparavant (par le locuteur ou par l'observateur lui-même, voir la deuxième condition), il y a identité de dénotation et on attribue aux deux expressions le même index de référent.
- Si aucune description complète n'est présente par ailleurs et s'il n'y a pas identité de référent avec une date dénotée par ailleurs, l'observateur note entre crochets la description complète de la date.

Les dates s'interprétant en fonction du moment de l'énonciation, on indique quel est ce moment en préambule et on lui attribue l'index de référent `o0`. Pour l'article analysé ici, on note :

<`[lundi 11 mai 1998, o0]`>

Quelques exemples, qui ont tous pour moment de l'énonciation la date ci-dessus :

(83) La Commission européenne a approuvé <vendredi `[8 mai 1998]`> l'acquisition des AGF par Allianz. Le feu bruxellois lève ainsi le dernier obstacle formel à la contre-offre publique d'achat (OPA) lancée en <février `[février 1998]`> sur l'assureur français.

(84) À <ce jour `[o0]`> l'accord de principe, qui devrait être scellé <le 20 mai `[20 mai 1998]`>, établit que, en échange des aides accordées au Lyonnais, les contreparties s'élèvent à 620 milliards de francs au moins.

(85) Il a reconquis la maîtrise des modalités de privatisation, à condition toutefois qu'elle intervienne d'<ici `[o0]`> à octobre 1999.

On notera dans le dernier exemple qu'il n'y a pas d'annotation pour l'expression *octobre 1999*, qui est une description complète du mois désigné par l'expression.

Toutes les références à une date ne sont pas nécessairement déictiques. Dans ce cas, elles relèvent du cas général, c'est-à-dire qu'on utilise de simples index de référent et une mise en relation des deux êtres par la relation `rel` :

(86) Pierre a rencontré Marie <le jour [o1]> de son anniversaire. <Le lendemain [o2]>, il l'a demandée en mariage.

2.7 Mise en perspective

Au terme de cette présentation de notre typologie des reprises [14], nous mettons notre approche en perspective avec celle qui est adoptée en sémantique formelle, puis nous comparons notre typologie avec l'analyse qu'ont proposée Halliday et Hasan des liens de cohésion en anglais [43], avec le schéma d'annotation proposé dans le projet MATE [59] et avec les description proposées récemment par S. Salmon-Alt [82, 81]. Les travaux sur les notions d'anaphore et de coréférence ou encore, à un niveau plus général, sur la notion d'identité sont extrêmement nombreux. Pour la mise en perspective de notre typologie, nous nous limitons à ceux qui sont décrits ici ; d'autres travaux seront décrits plus loin, sur la question de l'évaluation au chapitre 3 (section 3.7) et sur l'interprétation des expressions pronominales au chapitre 6.

2.7.1 Sémantique formelle

Pour décrire les phénomènes de reprises, nous avons fait usage de notions qui ne sont pas sans rappeler certaines notions utilisées dans la sémantique formelle, en particulier la notion d'univers de dénotation. Dans la mesure où la sémantique formelle constitue une approche classique en linguistique et probablement familière pour certains lecteurs, il nous semble utile de mettre en perspective notre approche avec celle qui est adoptée en sémantique formelle. Les différences sont en effet sensibles.

Nous rappelons ici les objectifs de la sémantique formelle, puis notre propres objectifs. Nous justifions ensuite notre choix de ne pas viser une représentation des textes dans un langage formel par notre volonté de traiter des textes effectifs. Nous revenons ensuite sur l'aspect méthodologique que nous défendons dans la thèse — nécessité de définir les conditions d'observations et de tester les hypothèses — en mettant à jour dans le même temps ce qui nous semble être certaines lacunes à cet égard dans les théories de sémantique formelle. Par contraste, notre approche s'en trouve justifiée.

Objectifs de la sémantique formelle. L'objectif de la sémantique formelle peut être résumé ainsi : il s'agit d'associer aux énoncés des « représentations sémantiques » qui sont des formules d'un langage logique quelconque (calcul des

[14] Dans la première partie de la thèse, les mises en perspective de notre travail par rapport à d'autres travaux (c'est-à-dire la présente section et les sections 3.7 et 4.9 dans les deux chapitres suivants) sont présentées après nos propositions. Ce choix de présentation n'implique nullement que nos propositions ont été définies sans connaissance des travaux évoqués.

prédicats, logique intensionnelle, etc.) à partir desquelles il sera possible d'effectuer des calculs. En général, le calcul visé inclut la détermination de la valeur de vérité des énoncés, ce qui passe par l'utilisation d'un univers de dénotation. Ces deux points sont résumés par la formule suivante, reprise de Chambreuil et Pariente [18, p. 64] et qui vise à caractériser l'objectif de la sémantique telle que pratiquée par R. Montague [86] :

> L'objectif principal de la sémantique est de caractériser les notions d'énoncé vrai et de conséquence.

La notion d'énoncé vrai est caractérisée par la mise en relation des formules logiques avec l'univers de dénotation. La notion de conséquence est caractérisée par les possibilités de calcul qu'offrent les formules logiques.

D'autres objectifs s'ajoutent à ces deux objectifs fondamentaux dans certaines théories. Ainsi la Discourse Representation Theory « est conçue comme une analyse (idéalisée) du processus par lequel le récepteur d'un énoncé arrive à saisir les pensées que cet énoncé contient [50, p. 8]. » Dans cette optique, les représentations sémantiques sont censées correspondre à la réalité de la pensée humaine : « le processus du raisonnement ne peut pas être compris autrement qu'en supposant que les croyances, désirs, etc. qui jouent le rôle des prémisses des inférences mentales et des conclusions qui en sont tirées ont une sorte de structure représentationnelle formelle, similaire à un langage... [50, p. 9]. »

Objectifs de la thèse. Nous rappelons ici les objectifs de la thèse et certains choix que nous avons fait :

a. nous voulons travailler sur des corpus de textes réels [15] ;
b. nous avons restreint notre champ d'observation à certains aspects de l'interprétation des textes : les phénomènes de reprises ;
c. l'objectif est d'implanter une machine qui traite effectivement des textes réels ;
d. l'objectif est d'attester que les conditions d'évaluation pour cette machine existent.

On voit ici que nos objectifs sont très différents de ceux de la sémantique formelle. Cette différence d'objectifs conduit à des différences dans la manière d'aborder les problèmes, différences que nous détaillons et justifions ici.

Textes réels et fragments de la langue. Notre choix de ne pas viser une traduction des énoncés dans des formules logiques découle de notre volonté de traiter des textes réels.

Dans les approches de sémantique formelle, on peut distinguer deux calculs :

– le calcul qu'on peut faire à partir d'un ensemble de formules logiques,

[15] Nous entendons par « textes réels » des textes qui ont été produits par des locuteurs indépendamment de toute visée illustrative du fonctionnement du langage.

– le calcul qui permet de passer des énoncés de la langue naturelle aux formules logiques.

Le premier type de calcul est en général spécifié avec le langage formel et c'est en grande partie ce qui fait l'intérêt de ce langage. C'est selon nous ce type de calcul qui permet de caractériser la notion de conséquence évoquée par Chambreuil et Pariente dans la citation donnée plus haut.

Le second type de calcul est plus problématique. En témoigne le fait que les différentes théories de sémantique formelle se limitent explicitement ou *de facto* à un fragment de la langue naturelle. À propos des travaux de R. Montague, Dowty et al. notent [30, p. 179] :

> La grammaire de PTQ, à laquelle Montague faisait référence comme un « fragment » de l'anglais, est délibérément restreinte pour les besoins de la présentation.

Dans le même ordre d'idée, Kamp et Reyle précisent quant à eux [50, p. 24] :

> Nous avons fait le choix d'une syntaxe qui associe à chacune des phrases du fragment de l'anglais que nous traiterons une structure syntaxique qui satisfait les besoins de la procédure d'interprétation que nous décrirons dans les chapitres suivants.

Les théories de sémantique formelle sont exposées sur la base d'un fragment de la langue et leur extension au-delà de ce fragment ne va pas toujours de soi ; Dowty et al., à la suite de la remarque citée ci-dessus, ajoutent :

> Dans beaucoup de cas, la manière dont le fragment de Montague pourrait être étendu pour traiter d'autres constructions de l'anglais est claire, mais dans beaucoup d'autres cas la possibilité d'un traitement simple et direct dans le cadre défini est problématique.

Notre objectif est de travailler sur des textes réels et la distance est très grande entre les fragments de la langue naturelle traités par les théories de sémantique formelle et les textes réels. C'est une des raisons pour lesquelles nous ne cherchons pas à associer aux énoncés des représentations sémantiques sous la forme de formules logiques.

Observations contrôlées. Si la portée des travaux de sémantique formelle est restreinte à un fragment du langage considéré, notre approche met aussi en jeu une vue fragmentaire de l'objet que nous étudions : nous nous limitons à l'observation des phénomènes de reprise. De telles limitations sont la règle dans une science du réel (ce qu'est la linguistique) :

> Dans une science du réel, on ne prétend pas décrire tout le réel. Celui-ci est modélisé, c'est-à-dire idéalisé de manière contrôlée pour l'observer selon le point de vue fixé. [10, p. 347]

Dans les deux premiers chapitres de la thèse, nous avons fixé le point de vue selon lequel nous voulons observer l'objet étudié et nous avons défini un langage de

notation des observations. Rappelons que ce langage de notation des observations vise à rendre compte du résultat de l'interprétation des expressions reliées entre elles et non du processus d'interprétation de ces expressions. On a là une différence avec les approches de sémantique formelle, où un des points clés est la traduction des énoncés de la langue naturelle dans les formules logiques.

La question de savoir si notre point de vue est meilleur ou moins bon que celui des travaux de sémantique formelle ne nous semble pas pertinente. Les deux approches offrent très certainement des vues complémentaires. Le point important ici est que nous avons effectivement explicité notre point de vue, c'est-à-dire définit les conditions dans lesquelles nous observerons le réel et évaluerons les hypothèses à venir sur ce réel.

Tester les hypothèses. L'intérêt du langage de notation des observations que nous avons défini est de permettre l'évaluation des hypothèses sur le processus d'interprétation des expressions qui seront susceptibles d'être formulées par la suite [16]. Ce langage permet en effet de comparer plusieurs interprétations d'un même texte effectuées, soit par des observateurs humains, soit par une machine. La comparaison d'observations faites par des observateurs différents permettra de mesurer l'inter-subjectivité de ces observations (cette problématique fait l'objet du chapitre 4). La comparaison d'une interprétation effectuée par une machine à l'interprétation d'un observateur humain, à supposer que cette dernière soit fiable, c'est-à-dire que l'inter-subjectivité des observations ait été établie, permettra d'évaluer les hypothèses implantées dans la machine.

La problématique est ici tout simplement celle des sciences du réel, que G. Bès, décrivant le paradigme 5P, caractérise ainsi [10, p. 347] :

> Le point méthodologique essentiel qui a guidé la structuration [du paradigme 5P] est l'objectif d'intégrer sérieusement la linguistique aux sciences du réel, dans lesquelles le progrès se fait par une interaction constante entre observation contrôlée, formulation d'hypothèses, calcul des conséquences de ces hypothèses, et test des conséquences obtenues par rapport à l'observé ou observable, aucune priorité temporelle ou causale n'étant prêtée à chacune de ces activités.

Dans cette même optique, l'extrait suivant apporte quelques précisions [11, p. 6] :

> Il serait facile de multiplier les citations pour montrer à quel point des expressions comme *théorie, hypothèse* et autres relevant de la méthodologie des sciences sont extraordinairement démonétisées en linguistique, et cela en linguistique standard ou classique et en lin-

[16] On rappelle que nous ne formulerons de telles hypothèses que pour un ensemble restreint d'expressions pronominales et non pour l'ensemble des phénomènes de reprise. Cela ne nous empêche pas de nous intéresser, en voyant au-delà de la présente thèse, au problème que représente la spécification de l'existence de conditions d'évaluation effectives pour les phénomènes de reprise.

> guistique formelle et/ou informatique. Dans ce document, *théorie* et *hypothèse* sont des expressions qui dénotent des formulations relevant d'une science du réel, ou de l'empirie, dans la terminologie de Granger [36], des formulations qui doivent être susceptibles d'être testées — ou qui doivent être falsifiables dans une terminologie popérienne — et qui, après test, peuvent être vérifiées ou non vérifiées [...]. Pour qu'une formulation acquière le statut d'hypothèse, et qu'un ensemble d'hypothèses acquière le statut de théorie, il faut que théorie et hypothèses soient associées à une manière précise de faire des observations et à une manière précise de comparer les déductions que l'on peu tirer des hypothèses avec les observations, représentées — elles aussi — d'une manière précise.

Nous voulons pratiquer la linguistique comme une science du réel et dans ce contexte, en reprenant les termes de la seconde citation de Bès, nous nous intéressons aux points suivants :

- les formulations que sont les théories et hypothèses *doivent être susceptibles d'être testées* ;
- théorie et hypothèses doivent être *associées à une manière précise de faire des observations et à une manière précise de comparer les déductions que l'on peu tirer des hypothèses avec les observations, représentées — elle aussi — d'une manière précise.*

Il n'y a dans la première partie de la présente thèse ni théorie, ni hypothèses. En revanche, nous abordons une question — celle de l'évaluation d'un système d'hypothèses ou d'une théorie — qui n'est précisément pas abordée dans les travaux dits « théoriques », parmi lesquels les travaux de sémantique formelle.

Quelles sont les conditions d'évaluation des théories de sémantique formelle ? La réponse n'est pas claire.

Nous avons vu qu'aussi bien les travaux de Montague que ceux de Kamp et Reyle se limitaient à un fragment de l'anglais. Les travaux en question doivent-ils être évalués au regard du seul fragment sur lequel ils portent ou sur l'ensemble de la langue à laquelle appartient le fragment ?

Dans leur présentation de la DRT, Kamp et Reyle réservent un chapitre aux problèmes non traités ou mal traités par leur théorie. Le chapitre en question est introduit comme suit [50, ch. 3, p. 233] :

> Dans ce chapitre, nous discutons un certain nombre de phénomènes ignorés dans les chapitres précédents, mais qui sont étroitement liés aux questions sur lesquelles nous avons porté notre attention. Ces phénomènes se répartissent en deux catégories. D'une part, les phénomènes qui affectent directement la théorie que nous avons présentée

> et qui, à proprement parler, l'invalident dans son état actuel [17]. Les autres phénomènes ne constituent pas une menace similaire pour le compte rendu donné jusqu'à présent [...].

L'usage de l'expression adverbiale *à proprement parler* suggère deux conditions d'évaluation possibles :

a. l'évaluation, que nous dirons « large », sur l'ensemble des phrases de l'anglais (qui invalide, *à proprement parler*, la théorie)
b. l'évaluation, que nous dirons « restreinte », sur le seul fragment de l'anglais à partir duquel ou pour lequel a été conçue la théorie.

L'existence de cette seconde possibilité d'évaluation est confirmée par la remarque suivante, faite par Kamp et Reyle à propos des réfléchis, dont la DRT rend mal compte [50, p. 234] :

> [La phrase *John admires himself*] ne fait pas partie du fragment de l'anglais que nous avons traité jusqu'à présent. Par conséquent, le fait que nous ne pouvons pas encore en rendre compte n'invalide pas nos propositions.

D'après cette remarque, il semble qu'il faille retenir l'évaluation « restreinte », c'est-à-dire considérer que la DRT doit être évaluée seulement au regard du fragment de l'anglais pour lequel elle a été définie.

Ce choix d'une évaluation restreinte au fragment dont la théorie est censé rendre compte est en fait le cas général en sémantique formelle. Il se manifeste dans le fait qu'on évalue les différentes théories au regard de la taille respective des fragments qu'elles couvrent (telle théorie traite ce phénomène, telle autre non). Ainsi, dans le projet FRACAS [18], pour comparer différentes approches de sémantique formelle, est spécifié « un petit fragment de l'anglais sous la forme d'une liste de phrases illustrant divers phénomènes. Le fragment est conçu pour couvrir des phénomènes dont l'intérêt est central dans les applications mettant en jeu de la sémantique computationnelle [49]. » Pour qui a l'objectif de travailler sur des textes réels, la question se pose de la représentativité du fragment considéré.

On peut par ailleurs s'interroger sur la pertinence de l'évaluation d'une théorie sémantique par comparaison avec une autre. Une évaluation de la théorie sémantique *en elle-même* (c'est-à-dire sans référence à d'autres théorie) devrait nous semble-t-il pouvoir être menée. Si une théorie sémantique doit être évaluée au regard du seul fragment sur lequel elle porte, alors une manière de l'évaluer serait peut-être de vérifier que tous les textes spécifiés par le composant d'analyse syntaxique du système (qui implique l'existence d'un lexique) reçoivent bien une (ou plusieurs) représentation(s) sémantique(s) qui correspondent avec la ou

[17] On the one hand there are those [phenomena] which directly affect the theory we have presented and which, strictly speaking, invalidate it as it stands.
[18] A Framework for Computational Semantics.

les interprétations qu'un observateur humain associerait à ces textes. Ce type d'évaluation, à supposer qu'il soit pertinent (ce que nous pensons être le cas), n'est pour ainsi dire jamais mené de manière systématique dans les travaux de sémantique formelle.

Pourtant, il serait susceptible de mettre à jour certaines inadéquations des formalismes proposés pour le fragment considéré lui-même. Nous illustrerons ce point par un exemple [19] d'inadéquation des représentations proposées par Montague pour un problème qui nous intéresse directement : la coréférence. Le problème est essentiellement que les représentations proposées ne permettent pas de spécifier l'absence de coréférence là où une telle absence doit être observée. Par exemple, la phrase :

(87) A soldier runs and a man sings.

est représentée par :

$$\exists x\ [\text{soldier'}(x) \wedge \text{run'}(x)] \wedge \exists y\ [\text{man'}(y) \wedge \text{sing'}(y)]$$

Rien n'empêche que la dénotation de x et de y dans cette formule soit la même, or un observateur effectif de l'exemple (87) n'envisagerait pas cette interprétation.

L'objectif de la présente section était de justifier les objectifs et les choix qui ont été fait dans la première partie de la thèse. Si nous l'avons fait en mettant en perspective notre approche avec les travaux de sémantique formelle, c'est parce ces travaux constituent pour beaucoup un cadre de référence. Nous avons justifié notre choix de ne pas viser une représentation de l'interprétation des textes sous la forme d'un langage logique par le fait que nous voulons traiter des textes réels et que les procédures de traduction des énoncés de la langue naturelle dans des formules logiques, telles qu'elles ont été définies à ce jour, nous paraissent trop incomplètes. Par ailleurs, nous avons justifié notre approche par le fait qu'elle intègre certaines exigences d'une science du réel, exigences dont ont a vu qu'elles étaient peu ou mal intégrées dans les travaux de sémantique formelle.

2.7.2 La cohésion selon Halliday et Hasan

M. Halliday et R. Hasan ont présenté dans un ouvrage qui fait figure de classique [43] une description des différents aspects qui font la cohésion d'un texte, celle-ci étant vu comme observable lorsque « l'interprétation d'un élément est dépendant de l'autre » [43, p. 4]. Notre typologie des liens de reprises décrit un ensemble de phénomènes plus restreints que ceux qui sont décrits par Halliday et Hasan, en particulier dans la mesure où ces derniers décrivent à la fois des relations exprimées explicitement par la structure de la phrase et des relations à distance. Notre typologie présente cependant certaines analogies avec le système de classification de Halliday et Hasan.

[19] Cet exemple nous a été signalé par G. Bès.

Dans leur description des différents liens possibles entre les éléments d'un texte, Halliday et Hasan distinguent les notions suivantes :

- la référence,
- la substitution et l'ellipse, cette dernière étant un cas particulier de substitution,
- la conjonction,
- la cohésion lexicale.

Les phénomènes décrits par la notion de conjonction ne relèvent pas de notre notion de reprise ; il s'agit essentiellement de l'usage de connecteurs tels que *et, en fait, cependant*, etc. Les notions de référence, substitution et cohésion lexicale, en revanche, sont pertinentes pour notre typologie.

La distinction entre référence et substitution est très proche de notre distinction entre des relations qui se situent au niveau de la dénotation et des relations qui se situent au niveau de la description [20] :

> Substitution is a relation between linguistic items, such as words or phrases ; whereas reference is a relation between meanings [43, p. 89].

Les phénomènes couverts par la notion de substitution chez Halliday et Hasan correspondent dans notre typologie à des identités de description ou à des reprises de type « paraphrase » (voir la section 2.5).

La notion de référence chez Halliday et Hasan est caractérisée essentiellement par un ensemble de formes : les pronoms, les articles définis et démonstratifs et les comparatifs, qui donnent lieu à trois types de référence (« personnelle », « démonstrative », « comparative »).

Les phénomènes relevant de la référence personnelle seront analysés par nous comme des identités de dénotation. Les phénomènes relevant de la référence démonstrative, dans la mesure où ils ne relèvent pas de la deixis, seront quant à eux analysés de manière plus disparate : soit une identité de dénotation, soit une relation `membre-de`, ou encore une relation référentielle. La notion de référence comparative est liée à l'emploi de certaines formes : les comparatifs, des adjectifs tels que *identique, similaire, même, différent* et *autre* ou certains adverbes qui leur correspondent. De manière générale, notre typologie ne rend pas compte de ce type de relation, si ce n'est par l'identité de description (on compare en général des objets de même type) ou, dans le cas de l'adjectif *autre*, par la relation `distingué-de`.

L'absence de correspondance directe entre les catégories de Halliday et Hasan et les nôtres résulte d'une différence essentielle : la classification de Halliday

[20] La terminologie de Halliday et Hasan étant relativement spécifique à leur étude, nous citons ici directement l'anglais. Le terme *meanings*, dans la terminologie de Halliday et Hasan, renvoie à la notion de dénotation. Voir, par exemple, la phrase : « The meaning of the reference item *he* is 'some person (male), other than the speaker or addressee, who can be identified by recourse to the environment' [43, p. 89]. »

et Hasan est basée principalement sur un ensemble de catégories lexicales ou grammaticales, alors que nous caractérisons les liens entre expressions en fonction du sens de la relation en jeu.

La notion de cohésion lexicale décrit « l'effet de cohésion atteint par le choix du vocabulaire ». Une partie des faits décrits sont des cas de coréférence [21], la notion de cohésion lexicale ayant dans ce cas surtout pour intérêt de caractériser des niveaux de généralité dans la classe des noms (en particulier l'identification de noms généraux, tels que *chose, personne, idée*, qui apparaissent souvent dans des syntagmes anaphoriques). Une autre partie des faits sont simplement des cas de reprise de description avec répétition. Les autres phénomènes tombant sous le concept de « cohésion lexicale » ont trait au fait que certains termes soient fréquemment co-occurrents dans les textes (un exemple est la paire *abeille* et *miel*) ; ces phénomènes dépassent le cadre de notre typologie des reprises. Dans une vision plus large, s'il y a une réelle dépendance entre des êtres désignés par des termes fréquemment co-occurents, cette dépendance devrait relever de notre relation `rel` générale.

Nous l'avons dit, Halliday et Hasan organisent leur description des relations de cohésion principalement en fonction de la forme des expressions. De ce fait, une même reprise pourra être analysée comme mettant en jeu plusieurs types de relations cohésives, en particulier dans le cas d'une reprise par un syntagme nominal défini avec identité de dénotation (p. ex. *une pomme*,... *le fruit*), où on aura à la fois une référence démonstrative et une cohésion lexicale. D'une certaine manière, il semble trivial que deux expressions qui désignent le même être mettent en jeu des termes qui sont sémantiquement liés. Pourtant, Halliday et Hasan ne cherchent pas à identifier de corrélations entre les différents types de liens qu'ils caractérisent. Un apport de notre typologie est selon nous l'identification d'une telle corrélation entre l'identité de description et les relations `membre-de` et `distingué-de`.

2.7.3 Le schéma MATE d'annotation de la coréférence

Le but du projet MATE (Multilevel Annotation Tools Engineering) est de développer des outils et des standards pour l'annotation de corpus. Un schéma d'annotation pour la coréférence au sens large a été défini [69], ainsi qu'un état de l'art des différents systèmes d'annotation des relations de coréférence et/ou des relations anaphoriques existants à la date du projet [28]. Le schéma d'annotation proposé dans MATE se veut un standard, qui intègre la plupart des relations visées dans les systèmes définis auparavant. C'est principalement avec ce standard

[21] Halliday et Hasan mentionnent ce point : « This might suggest that there was no distinct category of lexical cohesion ; that what we are calling 'lexical cohesion' was merely the reiteration of a lexical item in a context of grammatical cohesion, the cohesion being merely a matter of reference. »

que nous comparerons notre typologie et nous renvoyons le lecteur intéressé par les autres systèmes d'annotation à l'état de l'art fait dans le projet MATE [28].

Les relations envisagées dans le projet MATE sont les suivantes :

- « identité », lorsque « deux syntagmes nominaux dénotent le même objet du monde » ;
- deux relations « member » et « subset », décrivant respectivement la relation d'un élément ou d'un sous-ensemble à un ensemble ;
- trois relations de « possession » au sens large :
  - « attribut », p. ex. la taille ou le poids d'une personne est un attribut de cette personne ;
  - « partie », lorsqu'une expression dénote une partie physique de l'être dénoté par une autre expression, p. ex. les ailerons d'une fusée sont une partie de la fusée ;
  - « possession stricte », l'être dénoté par une expression appartient à l'être dénoté par une autre expression.
- « anaphores liées », relation réservée aux cas où une forme pronominale est liée par un antécédent qui est un quantificateur, comme, par exemple, *son* dans *Personne n'aimerait perdre son travail*.
- « fonction-valeur », p. ex. dans la phrase *La température est montée à 90 degrés, avant de redescendre à 70 degrés*, les syntagmes *90 degrés* et *70 degrés* sont vus comme dénotant chacun une valeur de la fonction dénotée par *La température* ;
- « instanciation », qui décrit la relation entre une expression qui dénote un être particulier d'une classe et l'expression qui dénote cette classe ;
- « relation événementielle », qui décrit le lien entre un être dénoté par un syntagme nominal et un événement ou une situation évoquée précédemment, p. ex. le lien entre *une explosion* et *le bruit* dans *Il y a eu une explosion ; le bruit était terrible* ;
- enfin, une relation générale pour couvrir les cas non décrits par les précédentes relations.

La relation d'identité de MATE correspond, du moins dans les termes de la définition, à notre notion de reprise avec identité de dénotation. On remarquera cependant que les cas d'« anaphore liée » dans MATE sont analysés par nous comme des cas de coréférence (voir la section « Modalités particulières dans les références aux ensembles », p. 47). La notion d'anaphore liée nous semble intégrer une dimension qui est de l'ordre de la syntaxe [22] et qui ne devrait pas selon nous entrer dans un schéma qui vise à spécifier l'interprétation des expressions et non le mécanisme d'interprétation des expressions. Par ailleurs, la manière dont sont

[22] Voir la définition qu'en donne T. Reinhart, citée dans [96, p. 91] : « L'anaphore liée requiert que l'antécédent c-commande l'anaphorique. Elle est illustrée par la relation entre un anaphorique et un antécédent quantifié. »

traités les cas où une expression anaphorique renvoie à un ensemble d'expression, que nous analysons comme relevant de la coréférence (voir p. 45 et p. 63), n'est pas spécifiée dans le projet MATE.

Il n'y a pas dans le projet MATE de notion qui corresponde à notre notion de « description », alors que celle-ci joue un rôle important dans notre typologie et a déjà été faite par ailleurs. Nous avons vu dans la section précédente que notre distinction entre dénotation et description peut être vue comme parallèle à celle que font Halliday et Hasan [43] entre « référence » et « substitution ». On notera que cette dernière a été utilisée dans le projet d'annotation Lancaster Anaphoric Treebank [33], ce qui rend son absence dans le projet MATE encore plus étonnante. Au final, à la lecture de [69], on ne sait pas comment annoter une phrase telle que :

(88) Wendy prefers the red T-shirt to the yellow one.

exemple pourtant mentionné en introduction comme un cas où la relation n'est pas l'identité de dénotation.

La notion de description nous permet par ailleurs de caractériser plus précisément les relations `membre-de` et `distingué-de`. On notera que rien ne ressemble à la relation `distingué-de` dans MATE et que notre relation `membre-de` subsume les trois relations ensemblistes de MATE : appartenance, inclusion et instanciation.

De manière générale, le projet MATE propose en effet des relations plus spécifiques que les nôtres : les relations « possession stricte », « attribut », « fonction-valeur » et « événementielle » ne sont pas caractérisées dans notre typologie. On notera que dans le cas des deux premières relations, tous les exemples donnés dans [69] mettent en jeu une dépendance structurelle à l'intérieur du syntagme nominal (p. ex. le lien entre *la voiture de Pierre* et *Pierre* est un lien de possession strict). De la même manière, les liens qui mettent en jeu la relation « fonction-valeur » sont pour nous explicites. Il nous semble que les liens anaphoriques (c'est-à-dire des liens à distance, non explicités par la syntaxe) qui mettent en jeu une de ces trois relations sont rares et il nous a semblé préférable de ne pas chercher à définir un ensemble de relations particulières qui, pour ne s'appliquer qu'à des cas très particuliers sans que ceux-ci aient été intégralement caractérisés, risquait d'être arbitraire.

2.7.4 L'annotation (co-)référentielle selon Salmon-Alt

Plus récemment, S. Salmon-Alt a proposé une description de l'interprétation des « expressions référentielles » [23] qui se traduit dans un nouveau schéma d'annotation de corpus [82, 81].

[23] Dans notre terminologie, les « expressions référentielles » de Salmon-Alt (comme celles des divers chercheurs du LORIA ou ayant travaillé autour du projet CERVICAL, voir [76] ou [72]) correspondent à des syntagmes nominaux dénotants.

Le schéma proposé par Salmon-Alt met en jeu trois relations :

- l'identité de dénotation (`ident`) ;
- la relation de « codomanialité » (`codom`) ;
- la relation d'« extraction » (`extract`).

Nous ne nous attarderons pas ici sur la relation `ident`, qui correspond à notre identité de dénotation, modulo le cas particulier des références à un ensemble évoqué plus loin. Nous décrivons les relations de codomanialité et d'extraction en reprenant les termes de Salmon-Alt :

- un lien de codomanialité s'établit entre deux entités discursives [24] $\mathtt{DE}_1$ et $\mathtt{DE}_2$ dont les référents $\mathtt{R}_1$ et $\mathtt{R}_2$ sont disjoints, à condition que l'identification de $\mathtt{R}_2$ se fasse par rapport à un objet-repère $\mathtt{R}_1$ [82, p. 213] ;
- un lien d'extraction s'établit entre deux entités discursives $\mathtt{DE}_1$ et $\mathtt{DE}_2$, lorsque le référent de $\mathtt{DE}_1$ est extrait du référent de $\mathtt{DE}_2$ [82, p. 214].

La notion de codomanialité elle-même renvoie à la notion de « domaine de référence », un domaine de référence étant « un ensemble contextuel local qui regroupe et structure des entités contextuelles de façon à prédire la distribution des différentes expressions référentielles [82, p. 119]. » Les « entités contextuelles » correspondent à peu près à des êtres de l'univers de dénotation dans notre terminologie.

La notion de domaine de référence découle d'une hypothèse que fait Salmon-Alt, qui est que « l'identification référentielle consiste en une opération de prélèvement d'un référent dans un ensemble [82, p. 104]. » Nous n'entrerons ici ni dans les différentes justifications qu'apporte Salmon-Alt à cette hypothèse, ni dans le « modèle » qui en découle. Ce dernier est censé [25] avoir une valeur explicative, en particulier de la distribution des différents types d'expressions référentielles (indéfinis, définis, pronoms, etc.).

Dans la mesure où notre approche vise à décrire l'interprétation d'un ensemble d'expressions apparaissant dans les textes de manière indépendante de toute hypothèse explicative, nous nous concentrerons plutôt sur la manière dont

[24] Le terme « entité discursive » est vague : dans le sens où il désigne des objets ayant un référent, il désigne des expressions, dans le sens où il s'oppose dans la terminologie de Salmon-Alt à « entité non discursive » (ou « entité de l'univers », c'est-à-dire les êtres qui sont identifiables dans la situation d'énonciation), il semble désigner les entités dénotées par les expressions. Peut-être faut-il y voir l'analogue des « discourse referents » de la DRT [50], c'est-à-dire quelque chose entre l'expression et l'être désigné par l'expression ; dans la DRT, à chaque expression est associé un *discourse referent* distinct et des prédicats d'identité lient ensuite ces *discourse referents* entre eux. On notera cependant que les *discourse referents* de la DRT sont des objets d'un langage formel, alors que ce n'est pas le cas des « entités discursives » de Salmon-Alt. Pour la présente discussion, nous considérerons le terme « entité discursive » comme équivalent à « expression référentielle » dans la terminologie de Salmon-Alt.

[25] L'hypothèse de Salmon-Alt n'est pas à proprement parler *testée* par elle. L'hypothèse en question est justifiée à partir d'exemples et de description des données linguistiques divers, mais elle n'est pas évaluée au regard de données qui n'ont pas servi à son élaboration.

les descriptions de Salmon-Alt se traduisent dans son schéma d'annotation de corpus.

Les liens de codomanialité sont utilisés par Salmon-Alt pour rendre compte des relations suivantes dans notre terminologie [26] :

a. identité de description entre deux expressions dénotantes (voir la sous-section 2.2.1) ;
b. relation `distingué-de` (voir la section 2.4) ;
c. relation entre expressions dont les référents constituent un ensemble auquel il est fait référence par la suite (cas dont nous rendons compte comme étant un type particulier d'identité de dénotation, voir la sous-section « Reprise dont la source est un ensemble d'expressions », page 45).

On notera que toutes ces relations sont des relations de reprise qui mettent en jeu une identité de description [27].

Notre notion d'identité de description recouvre cependant un ensemble de phénomènes plus larges que ceux qui sont décrits par les liens de codomanialité. En faisant abstraction des reprises mettant en jeu une expression non dénotante (voir la sous-section 2.2.2), dont Salmon-Alt ne parle pas mais qu'on peut considérer comme implicitement exclues de son champ d'observation par le fait qu'elles ne mettent pas en jeu des expression référentielles, l'identité de description se retrouve dans les liens qui mettent en jeu dans notre terminologie une relation `membre-de`.

Dans la terminologie de Salmon-Alt, les reprises avec relation `membre-de` sont décrites de la même manière que les phénomènes que nous regroupons sous la notion de « relations référentielles » (voir la sous-section 2.6.1), c'est-à-dire comme mettant en jeu un lien d'« extraction ». La notion d'extraction s'entend en relation avec celle de domaine de référence et de partition d'un domaine de référence : « l'interprétation des « anaphores associatives » s'appuie sur des connaissances sur une éventuelle décomposition — c'est-à-dire partition — d'un domaine donné [82, p. 120]. »

Considérons les deux exemples suivants, repris de [82, p. 125] :

(89) Il faut faire des pyramides. Voilà la première.

(90) Je vais faire la maison. Il faut mettre un toit dessus.

[26] Ces trois configurations sont exposées explicitement dans [81].

[27] Salmon-Alt dit dans [81, section « Conclusion »] que son schéma d'annotation « devrait permettre d'annoter, dans un cadre cohérent, un certain nombre de phénomènes jusqu'alors traités par des conventions *ad hoc* (telles que le lien « description » défini par [90]) ». Comme nous avons été partie prenante dans le projet décrit dans [90], nous nous permettons de répondre ici à cette critique : les liens de codomanialité de Salmon-Alt spécifient des ensembles (ou domaines) qui sont toujours caractérisés par une identité du type des référents du domaine, c'est-à-dire par une description. Les notions d'identité de description et de codomanialité sont donc très proches et il n'y a pas lieu de qualifier l'une de *ad hoc* (avec un sens péjoratif) par opposition à l'autre.

En (89), l'interprétation de *la première* est décrite comme suit :

- à l'expression *des pyramides* est associée une « représentation mentale », qui constitue le domaine de référence d de l'expression *la première* ;
- l'identification du référent de *la première* passe par une partition du domaine de référence d et identification d'un élément de cette partition, le référent de *la première*.

L'interprétation de *un toit* dans (90) est décrite de manière similaire. Le domaine de référence d de l'expression *un toit* est ici la représentation mentale associée à *une maison* et l'identification du référent de *un toit* passe par une partition du domaine de référence d. Un domaine de référence, dans le système descriptif de Salmon-Alt, peut toujours être vu comme un ensemble d'entités contextuelles. C'est le cas ici pour la représentation mentale associée à *une maison*.

Les deux exemples (89) et (90) doivent être décrits dans notre système comme mettant en jeu respectivement une relation `membre-de` et une relation `partie-de`. Ces deux relations relèvent d'une même notion d'« extraction » chez Salmon-Alt, mais il y a pour nous une différence notable entre les deux, dans le sens où l'une met en jeu une identité de description et l'autre non (voir la discussion de ce point p. 73). On notera que de cette distinction découle un des points importants de notre typologie des reprises, qui est de restreindre la notion d'ensemble à des ensembles d'êtres *de même type*, alors que dans le système descriptif de Salmon-Alt, tout, y compris ce que nous appelons les « êtres singuliers » (voir section 1.4.4), peut être vu comme un ensemble, ce qui risque de rendre les notions d'ensemble ou de domaine de référence peu opérationnelles.

Par ailleurs, à partir des résultats de l'expérience décrite au chapitre 4, on peut faire l'hypothèse que les liens de type `membre-de` seront plus susceptibles d'être observés de manière inter-subjective. Si cette hypothèse se trouvait confirmée, on aurait là une justification supplémentaire de notre notion de reprise.

On pourra être tenté de juger de l'intérêt ou de la pertinence de notre système descriptif par rapport à celui que propose S. Salmon-Alt, ou inversement. Il importe cependant de noter que cela ne pourra être fait qu'en tenant compte des objectifs respectifs des deux systèmes, qui sont très différents.

Nous voulons décrire un ensemble de relations entre expressions qui ne soit pas restreintes par une spécification par la forme des expressions et qui soient observables de manière inter-subjective. Salmon-Alt, pour sa part, ne met à aucun moment en question l'opérationalité des notions qu'elle utilise, dans le sens d'une interrogation sur le fait que celles-ci puissent être manipulées de manière inter-subjective par des observateurs différents.

Chez Salmon-Alt, l'hypothèse d'une identification référentielle passant par une opération d'extraction (évoquée plus haut) traduit une volonté de proposer un modèle du processus cognitif qui conduit à l'interprétation des expressions

référentielles. Nous avons pour notre part, dans notre description des données, voulu autant que possible faire abstraction de ce qui relevait du processus d'interprétation (voir section 1.3.1).

À elles seules, ces différences d'objectifs suffisent selon nous à expliquer les divergences entre Salmon-Alt et nous dans le découpage des données linguistiques.

Cela étant, il nous semble que les différentes approches décrites dans cette section 2.7 ne sont pas exclusives l'une de l'autre. Elles reflètent le fait qu'il y a toujours différentes manières de voir un même objet. Ce serait selon nous une erreur de penser qu'il n'y en a qu'une qui soit LA bonne manière de voir.

Chapitre 3

Critères d'évaluation

Nous avons présenté au chapitre précédent une typologie des reprises. Dans la perspective, d'une part, de tester l'opérationalité de la typologie présentée, d'autre part, de poursuivre la description des phénomènes de reprise jusqu'à la formulation d'hypothèses sur ce qui régit ces phénomènes, il nous faut spécifier des critères d'évaluation [1].

Nous proposerons des critères d'évaluation pour les différents types de reprises que nous avons définis, mais le présent chapitre sera limité à la description du système d'évaluation que nous proposons pour l'évaluation de l'identification des reprises avec identité de dénotation, ou coréférence. La présentation des critères d'évaluation pour les reprises avec identité de dénotation, sujet qui selon nous soulève le plus de questions pour l'évaluation, nous donnera tous les éléments pour définir ensuite les critères d'évaluation pour les autres types de reprises (relations `membre-de`, `distingué-de`, etc.), critères qui seront présentés dans le chapitre suivant.

Après une présentation de la problématique de l'évaluation, son intérêt et les données en jeu (section 3.1), nous proposons un critère d'évaluation pour la résolution des coréférences en termes d'« assignation de dénotation ». Les chaînes de coréférence sont définies par la propriété qu'ont les expressions qu'elles contiennent de dénoter un référent particulier et le but est alors d'évaluer si le bon référent a été associé aux expressions (section 3.2).

Étant donné une analyse supposée correcte des chaînes de coréférence d'un texte et une seconde analyse à évaluer au regard de cette version correcte, notre critère d'évaluation requiert une mise en correspondance des référents identifiés dans chacune des analyses. La section 3.3 décrit les contraintes qui régissent

[1] Le travail présenté dans ce chapitre a fait l'objet d'une publication [89]. Nous signalons au lecteur intéressé par la problématique abordée ici un article de Kehler et al. [52], dont le propos n'est pas sans relation avec certaines idées développées dans ce chapitre. L'article en question a paru bien après l'accomplissement du travail décrit ici et les contraintes de temps ne nous ont pas permis d'en faire état de manière détaillée.

cette correspondance ; en particulier, nous observons que le fait que les expressions appartenant à une chaîne de coréférence soient plus ou moins spécifiques relativement à leur référent commun doit être pris en compte.

La section suivante (3.4) présente la méthode que nous avons définie pour implanter le critère d'évaluation proposé et les contraintes qui en dérivent. Le fonctionnement de cette méthode est illustré par un exemple (section 3.5).

Une fois que la correspondance entre les référents des deux analyses est établie, les assignations de dénotation de la réponse peuvent être comparées à celles de l'annotation clé. À partir de cette comparaison, nous calculons des mesures d'évaluation. Outre les mesures classiques de rappel et précision, nous proposons des mesures supplémentaires pour une analyse plus fine des erreurs (section 3.6).

Enfin, après avoir présenté dans le détail notre méthode d'évaluation pour la coréférence, nous la comparons dans la dernière section (3.7) avec trois méthodes existantes : le système développé par Vilain et al. [91] pour le *Coreference Task* de MUC-6 [44], la méthode « classes noyaux exclusifs » de Popescu-Belis [71, 72, 74] et enfin « l'algorithme B-3 » de Bagga et Baldwin [5].

3.1 Problématique

3.1.1 Intérêt des critères d'évaluation

Tester l'opérationalité des descriptions linguistiques.

Le chapitre suivant décrit une expérience visant à attester que l'observation des liens de reprise et des relations qui peuvent être observées dans les cas d'anaphore associative est inter-subjective, c'est-à-dire que différents observateurs font bien les mêmes observations. Plus précisément, nous chercherons à savoir si des personnes différentes font bien les mêmes observations que nous : les observations doivent être inter-subjectives et l'être dans le respect des définitions que nous avons posées, cette seconde condition caractérisant l'opérationalité de notre typologie. Quelques étudiants du GRIL ont donc été invités à nous dire quelles reprises (ou autres relations entre expressions) ils observaient dans quelques textes. Il nous faut donner les critères qui nous permettrons de dire dans quelle mesure ces observations sont correctes.

On notera que si notre typologie ne se révèle pas opérationnelle, alors les conditions d'évaluation d'un éventuel système d'identification automatique des liens de reprises n'existeront pas.

Valider les hypothèses de règles.

Dans la perspective du développement d'un système d'interprétation automatique des textes, nous serons amenés à formuler des règles visant à décrire plus avant les phénomènes de reprises. Il nous faudra évaluer la validité de ces règles

à travers une évaluation des résultats obtenus par ce programme. Les critères d'évaluation présentés ici serviront également cet objectif [2].

3.1.2 Données de l'évaluation

Prédicats d'observation

Les objets dont la qualité est à évaluer sont dans notre cas des observations faites sur des textes, un ensemble de commentaires sur ces textes. Par exemple, dans notre schéma d'annotation, annoter deux expressions avec le même index de référent, c'est dire « ces deux expressions ont la même dénotation ». Sémantiquement, c'est donc la validité d'un discours sur un texte que nous souhaitons évaluer. Ce discours est constitué d'un ensemble de prédicats d'observation, un prédicat d'observation étant, dans notre cas, une formule telle que « je vois une reprise de type x entre telle expression et telle autre expression ». Concrètement, les prédicats d'observations sont exprimés sous la forme d'annotations insérées dans le texte analysé. Étant donné le texte suivant, par exemple,

(1) Pour la CJCE, l'objet d'une convention n'est pas de garantir au contribuable que l'imposition due dans un État ne soit pas supérieure à celle qu'il doit payer dans l'autre.

l'annotation suivante exprime trois observations : une reprise avec identité de dénotation entre *il* et *au contribuable*, une reprise de description entre le syntagme dont le noyau est *celle* et le syntagme dont le noyau est *imposition*, et une reprise avec relation `distingué-de` entre *l'autre* et *un État*.

(2) Pour la CJCE, l'objet d'une convention n'est pas de garantir au <contribuable `[o1]`> que <l'imposition `[imposition]`> due dans <un État `[o2]`> ne soit pas supérieure à <celle `[imposition]`> qu'<il `[o1]`> doit payer dans <l'autre `[o3, o3-dde-o2]`>.

Annotation de référence et annotation réponse

Pour évaluer la qualité d'un ensemble de prédicats d'observations, il faut les comparer à un étalon représentant la qualité optimale. Cet étalon sera pour nous l'ensemble des prédicats d'observations que nous-mêmes formulerons sous la forme de ce qui sera alors l'« annotation de référence ». Pour ce qui concerne les reprises, l'annotation proposée ci-dessus en (2) serait susceptible, dans la mesure où on la considère complète et correcte, de constituer une annotation de référence pour le texte en (1).

On peut voir l'identification et l'interprétation des reprises dans les textes comme le problème que nous soumettons à une résolution par un observateur ou

[2] Sur la nécessité d'attester l'inter-subjectivité des observations et de tester les hypothèses, voir [10, p. 285].

un programme informatique ; l'annotation proposée par l'observateur ou la machine — qui exprime un ensemble d'observations — constituera une réponse à ce problème, réponse que nous évaluerons au regard de la solution idéale constituée par l'annotation de référence. Quelle que soit la source ayant produite l'annotation que l'on souhaite évaluer, la problématique sera donc de comparer cette annotation avec l'annotation de référence contenant la solution au problème posé et de mesurer les écarts entre les deux. Suivant la terminologie habituelle, nous appellerons l'annotation de référence la « clé » ; l'annotation à évaluer sera appelée la « réponse ».

Prédicats d'évaluation

Évaluer une annotation réponse au regard d'une annotation de référence, c'est formuler un jugement sur les prédicats d'observation exprimés par l'annotation réponse. De manière minimale, il s'agit de dire si les prédicats d'observation de la réponse sont corrects ou incorrects. Nous appelons de tels jugements, qui portent un jugement de valeur sur des prédicats d'observation, des prédicats d'évaluation.

Les jugements de valeur exprimés par les prédicats d'évaluation pourront être variés, et non pas simplement se limiter à une dualité correct/incorrect. Supposons qu'on veuille, par exemple, évaluer l'annotation suivante du texte présenté en (1), au regard de l'annotation de référence présentée en (2) :

(3) Pour la CJCE, l'objet d'une convention n'est pas de garantir au contribuable que <l'imposition `[imposition]`> due dans <un État `[o1]`> ne soit pas supérieure à <celle `[imposition]`> qu'<il `[o1]`> doit payer dans l'autre.

Cette annotation exprime deux observations : une reprise avec identité de dénotation entre *il* et *un État* et une reprise de description entre le syntagme dont le noyau est *celle* et le syntagme dont le noyau est *imposition*. Une manière de juger cette annotation est de dire qu'elle contient deux observations dont l'une est correcte et l'autre incorrecte. On peut aussi noter que la coréférence entre *il* et *un État* dans l'annotation réponse correspond d'une certaine façon à la coréférence entre *il* et *au contribuable* dans l'annotation clé — le problème étant l'interprétation du pronom *il*. Enfin, et on pourra vouloir considérer que cette erreur n'est pas du même type que la précédente, l'annotation clé contient une observation que n'exprime pas l'annotation réponse. Nous dirons donc qu'une observation de la clé est *manquante* dans la réponse et que la réponse contient une observation *correcte* et une observation *incorrecte*. Ces trois jugements sont autant de prédicats d'évaluation.

Mesures d'évaluation

Pour mesurer la qualité d'une analyse d'un ensemble de données par rapport à une autre analyse, on utilise couramment en linguistique informatique deux

mesures d'évaluation : le « rappel » et « la précision ». Si on considère que l'annotation clé contient un ensemble d'observations, une réponse est parfaite si elle contient toutes et seulement les observations qui sont présentes dans la clé. Dans le cas où la réponse n'est pas parfaite, la mesure du rappel vise à évaluer la proportion d'observations correctes dans cette réponse par rapport aux observations de l'annotation clé. La mesure de précision, pour sa part, vise à évaluer la proportion d'observations correctes dans la réponse par rapport à l'ensemble des observations que cette réponse contient.

Étant donné les valeurs suivantes :

possible = nombre d'observations dans la clé
effectif = nombre d'observations dans la réponse
correct = nombre d'observations correctes dans la réponse

rappel et précision correspondent alors aux rapports suivants :

rappel = *correct*/*possible*
précision = *correct*/*effectif*

Nous utiliserons ces mesures pour évaluer la qualité des observations faites par un observateur ou une machine au regard des observations que nous-mêmes faisons. On remarque que la notion d'« observation correcte », qui relève des prédicats d'évaluation, joue un rôle central.

Pour l'annotation présentée en (3), évaluée au regard de l'annotation de référence présentée en (2) avec les prédicats d'évaluation proposés ci-dessus, nous aurions un rappel de 1/3 (une observation correcte dans la réponse sur trois possibles dans la clé) et une précision de 1/2 (une observation correcte dans la réponse sur deux observations effectivement présentes dans la réponse), *a priori*, une pas très bonne note.

Les exemples proposés ici ne sont destinés qu'à donner une première approche de la problématique de l'évaluation. Après cette vue générale, nous en venons au point central de ce chapitre : les critères et la méthode d'évaluation que nous proposons pour la résolution des coréférences.

3.2 Évaluer l'assignation des dénotations

La distinction principale entre description et dénotation dans notre typologie conduit l'observateur à faire des observations qui portent tantôt sur des expressions, tantôt sur la *dénotation* des expressions. Comparer des observations portant sur des expressions ne devrait pas poser de problème majeur, dans la mesure où les expressions ont une réalité concrète (dans notre cas elles sont des suites de caractères). Par contre, comparer des observations portant sur la dénotation des expressions s'avère plus problématique, dans la mesure où les référents des expressions sont des objets qui sont externes au texte, qui n'existent que dans la

mesure où un observateur les identifie. Le critère d'évaluation que nous présentons ici pour les reprises avec identité de dénotation vise à prendre en compte le fait que l'identification de ces reprises met en jeu l'identification des référents des expressions et non seulement les expressions.

On appelle « chaîne de référence » l'ensemble des expressions qui, dans un texte, dénotent un même référent. Étant donné un texte T, la relation entre les chaînes de référence et les référents est telle que pour chaque chaîne de référence CR, il existe un référent unique R, tel que :

$$CR = \{x \mid x \text{ est une expression qui dénote } R \text{ dans } T\}$$

Une chaîne de référence, selon notre définition, peut être un singleton. Dans la phrase *Pierre aime Marie*, par exemple, le singleton $\{Pierre\}$, c'est-à-dire l'ensemble des expressions qui dénotent Pierre dans ce texte, est une chaîne de référence.

Identifier les reprises de dénotation avec identité dans un texte donné revient à identifier des chaînes de référence qui contiennent au moins deux éléments. Nous appelons de telles chaînes de référence des « chaînes de coréférence ».

Considérons le texte présenté figure 3.1, extrait d'un article du Monde datant approximativement de la fin de 1986 - début 1987. Le contexte est celui de la cohabitation entre François Mitterrand et Jacques Chirac. On a annoté dans cet extrait les expressions entrant dans une relation de reprise avec identité de dénotation [3].

> En apportant < son$_{(1)}$ `[o1]`> appui au principe de la lutte contre < l'inflation$_{(1)}$ `[o2]`>, quatre jours après avoir engagé le débat sur le thème du dialogue social, < M. François Mitterrand `[o1]`> a donné la preuve que la position du < gouvernement$_{(1)}$ `[o3]`>, expliquée depuis la fin de la semaine dernière par M. Chirac, n'est pas facile à attaquer. Certes, < le président de la République `[o1]`> ne renie nullement < son$_{(2)}$ `[o1]`> geste du 1er janvier lorsqu'< il$_{(1)}$ `[o1]`> avait reçu des représentants des cheminots en grève, et < il$_{(2)}$ `[o1]`> maintient < sa `[o1]`> critique d'une rigueur inégalement partagée. Il$_{(3)}$ reste que, parti sur la défense de la « cohésion sociale », < il$_{(4)}$ `[o1]`> a jugé prudent de < s' `[o1]`>affirmer « en phase » avec < le gouvernement$_{(2)}$ `[o3]`>, pour une fois, sur la fermeté face au risque d'une relance de < l'inflation$_{(2)}$ `[o2]`>.

FIG. 3.1 – Extrait 1. Annotation clé.

Ce texte contient trois chaînes de coréférence. De manière générale, nous utiliserons, pour désigner une chaîne de référence, une séquence de la forme A_{o_i}, où A est une lettre capitale et o_i un identifiant de référent. Pour le texte donné en

[3] Les différentes occurrences d'expressions identiques sont différenciées par des indices.

exemple, nous désignerons les trois chaînes de coréférence par C_{o1}, C_{o2} et C_{o3} (C pour « clé ») :

C_{o1} = { *son*$_{(1)}$, *M. François Mitterrand, le président de la République,* *son*$_{(2)}$, *il*$_{(1)}$, *il*$_{(2)}$, *sa*, *il*$_{(4)}$, *s'* }
C_{o2} = { *l'inflation*$_{(1)}$, *l'inflation*$_{(2)}$ }
C_{o3} = { *le gouvernement*$_{(1)}$, *le gouvernement*$_{(2)}$ }

On notera que le texte contient par ailleurs un certain nombre de chaînes de référence à un élément, par exemple les trois chaînes suivantes, notées C_{o4}, C_{o5} et C_{o6} par convention :

C_{o4} = { *la lutte contre l'inflation* }
C_{o5} = { *le dialogue social* }
C_{o6} = { *M. Chirac* }

Le critère d'évaluation que nous posons pour l'identification des reprises avec identité de dénotation vise à tenir compte du fait que l'identification des chaînes de coréférence met en jeu une relation entre expressions basée sur la *dénotation* de ces expressions. Comme l'indique la définition des chaînes de référence proposée ci-dessus, ce qui caractérise une chaîne de référence, c'est la propriété qu'ont les expressions qui appartiennent à cette chaîne de dénoter un référent particulier. Les trois chaînes de coréférence de notre exemple (figure 3.1) sont ainsi caractérisées par les treize affirmations suivantes :

- *son*$_{(1)}$ dénote $o1$,
- *M. François Mitterrand* dénote $o1$,
- *le président de la République* dénote $o1$,
- *son*$_{(2)}$ dénote $o1$,
- *il*$_{(1)}$ dénote $o1$,
- *il*$_{(2)}$ dénote $o1$,
- *sa* dénote $o1$,
- *il*$_{(4)}$ dénote $o1$,
- *s'* dénote $o1$,
- *l'inflation*$_{(1)}$ dénote $o2$,
- *l'inflation*$_{(2)}$ dénote $o2$,
- *le gouvernement*$_{(1)}$ dénote $o3$,
- *le gouvernement*$_{(2)}$ dénote $o3$.

Considérant que le point crucial dans la résolution des coréférences est la propriété qu'ont les expressions de dénoter un référent particulier, nous proposons un système d'évaluation dont le but sera d'évaluer si la dénotation assignée aux expressions est correcte ou non. L'identification des chaînes de coréférence induit un ensemble de prédicats d'observation de la forme :

- l'expression e_i dénote le référent o_i.

Nous appellerons de tels prédicats des « assignations de dénotation ». Notre but sera d'évaluer si les assignations de dénotation formulées dans une annotation sont correctes ou non par rapport à celles qui sont formulées dans une annotation de référence.

Nous pensons que faire le lien entre les expressions et leurs référents, c'est-à-dire assigner une dénotation aux expressions, est le but de l'identification des chaînes de coréférence, résultat à partir duquel on pense être mieux à même d'extraire du texte les différentes informations données pour chaque référent.

3.3 Correspondance entre clé et réponse

La comparaison de nos deux annotations, l'annotation de référence, ou « clé », d'une part, et l'annotation à évaluer, ou « réponse », d'autre part, sera donc une comparaison de deux ensembles d'assignations de dénotation. Nous dirons d'un prédicat d'observation « e_1 dénote o_1 » dans la sortie est correct si la clé contient le prédicat d'observation « e_1 dénote o_2 » et o_1 et o_2 désignent le même référent. Le problème est maintenant de savoir si o_1 et o_2 sont bien les mêmes référents : étant donné un ensemble de référents dans la clé et un ensemble de référents dans la réponse, nous devons dire quel référent de la clé correspond à quel référent de la réponse, et réciproquement.

Pour déterminer la correspondance entre les référents de la clé et ceux de la réponse, nous profiterons du fait que l'origine des référents se situe dans le discours lui-même. Avant d'en venir à ce point, nous pouvons cependant faire une remarque sur la nature de la correspondance recherchée.

3.3.1 Correspondance de type 1-1

À partir du moment où le critère d'évaluation est fixé comme une évaluation des assignations de dénotation, on peut faire l'observation suivante :

> Étant donné e_1 et e_2 deux expressions d'un texte T et étant donné A et B deux analyses des chaînes de coréférence de T, si e_1 et e_2 appartiennent à une même chaîne de référence (en l'occurrence une chaîne de coréférence) dans une annotation et si e_1 et e_2 appartiennent à deux chaînes de référence distinctes (quel qu'en soit le type) dans l'autre annotation, alors e_1 et e_2 ne peuvent avoir toutes deux reçu une dénotation correcte dans les deux annotations.

Supposons deux annotations A et B d'un même texte. Mettons que l'annotation A contienne la chaîne de coréférence suivante, les chiffres faisant référence à des expressions :

$$A_{o1} = \{1, 2, 3, 4, 5, 6, 7\}$$

et que les expressions de A_{o1} se trouvent appartenir à trois chaînes de référence distinctes dans l'annotation B :

$$B_{o1'} = \{1, 2\}$$
$$B_{o2'} = \{3, 4\}$$
$$B_{o3'} = \{5, 6, 7\}$$

Si l'annotation clé est A, alors les sept expressions ont la propriété commune de dénoter un référent particulier. Parmi les assignations de dénotation effectuées dans l'annotation B (la réponse), nous avons :

- 1 dénote $o1'$
- 3 dénote $o2'$
- 5 dénote $o3'$

Si l'on considère l'un de ces prédicats d'observation correct, alors on doit considérer que les deux autres sont incorrects. De manière générale, dans ce cas, les assignations de dénotation ne peuvent être correctes que pour les expressions d'une et une seule chaîne de référence de B. En d'autres termes, seul l'un des trois référents identifiés dans B ($o1'$, $o2'$ ou $o3'$) peut correspondre au référent identifié dans A ($o1$).

Inversement, si l'annotation clé est B, les assignations de dénotation de A ne peuvent être correctes que pour les expressions qui appartiennent à un et un seul des sous-ensembles de A_{o1} qui correspondent respectivement à $B_{o1'}$, $B_{o2'}$ et $B_{o3'}$. En d'autres termes le référent o_1 identifié dans A correspond soit à $o1'$, soit à $o2'$, soit à $o3'$ et à un seul de ces trois référents.

En résumé, notre critère d'évaluation requiert qu'il y ait une correspondance de type 1-1 entre les référents de l'annotation clé et ceux de la réponse.

Pour revenir sur l'exemple qui nous a servi à illustrer notre propos, nous considérerons donc, comme le fait Popescu-Belis [74] dans sa méthode « classes-noyaux-exclusifs » [4] et contrairement à la méthode proposée par Vilain et al. [91], que dans le cas où A est l'annotation clé, la réponse identifie trois référents là où il n'y en a qu'un et que les expressions de deux des trois chaînes de référence de B n'ont pas reçu la bonne dénotation. Dans le cas où la clé est B, la réponse identifie un seul référent là où elle aurait dû en identifier trois ; les expressions qui auraient dû être reliées aux deux référents non identifiés n'ont pas reçu la bonne dénotation.

3.3.2 Spécificité des descriptions

Jusqu'à présent, nous avons parlé des référents des expressions comme si nous avions effectivement accès à ces objets dans les annotations à comparer, mais les

[4] « Exclusive Core-MRs » dans l'article cité. La traduction est tiré d'un article postérieur de Popescu-Belis [71, 72].

référents n'existent que dans la mesure où un observateur humain est là pour dire qu'ils existent. Étant donné un texte, l'observateur humain associe des référents aux expressions et identifie ainsi des chaînes de référence, mais il ne nous reste que le résultat du processus, à savoir des ensembles d'expressions. L'observateur a bien différencié les différents référents qu'il associait aux expressions au moyen d'indices de la forme o_i, mais ces indices ne sont nullement des constantes qui identifieraient de manière rigide un référent à travers différent textes, en particulier à travers l'annotation clé et l'annotation réponse. Pour établir la correspondance entre clé et réponse, ces indices doivent donc à leur tour être interprétés, dans le sens où « interpréter » veut dire « associer un référent ». Le problème est en quelque sorte d'établir les liens de coréférence entre les chaînes de référence de la clé et celles de la réponse. Pour nos deux annotations A et B ci-dessus, il nous faut dire qui de $B_{o1'}$, $B_{o2'}$ ou $B_{o3'}$ dénote le même référent que A_{o1}.

Pour répondre à ce problème, nous nous plaçons dans une situation où le résultat d'une identification des chaînes de références dans un texte serait interprété par un observateur humain ; autrement dit, étant donné un ensemble de chaînes de références, nous envisageons le cas où un observateur humain aurait à associer un référent à ces chaînes.

Supposons qu'on prenne une chaîne de coréférence et qu'on demande à un observateur humain de dire à quoi elle fait référence, par exemple la chaîne :

$$A_{o1} = \{\textit{Bill Clinton, son, le président des États-Unis, il}\}$$

Il est plus que probable que cet observateur reconnaîtra comme référent de cette chaîne de référence l'homme dont le second mandat de président des États-Unis s'est terminé en janvier de l'année 2001.

La remarque que nous faisons à partir de cet exemple est que les expressions dans la chaîne de référence aurons contribué à des degrés divers à l'identification de son référent. L'expression *Bill Clinton*, en elle-même, aurait suffit pour identifier le référent, alors que l'identification du référent serait impossible sur la seule base des expressions *le président des États-Unis*, *son* et *il* : pour être interprétées, il faut que ces expressions soient mises en relation avec leur contexte par une relation anaphorique.

Pour prendre un autre exemple, supposons qu'un observateur quelconque ait identifié dans le texte présenté figure 3.1 la chaîne de référence suivante :

$$R_{o1'} = \{\ \textit{M. Chirac, le président de la République, } son_{(2)},\ il_{(1)},\ il_{(2)},\ sa,\ il_{(4)},\ s'\ \}$$

Étant donné cette chaîne de référence, un observateur serait, selon nous, conduit à penser qu'elle dénote Jacques Chirac, tandis que les expressions de la chaîne C_{o1} dénotent François Mitterrand. Nous dirons donc que les expressions *le président de la République,* $son_{(2)}$, $il_{(1)}$, $il_{(2)}$, *sa,* $il_{(4)}$ et *s'* n'ont pas reçu la bonne dénotation.

Il est important de noter qu'on ne considère pas que l'expression *M. Chirac* n'a pas reçu la bonne dénotation. Nous dirons plutôt que la chaîne $R_{o1'}$ correspond à la chaîne $C_{o6} = \{M.\ Chirac\}$ de l'annotation clé.

Pour associer un référent à une chaîne de référence, il faut interpréter au moins une expression de cette chaîne. Nous considérons que, pour cette expression, l'assignation de dénotation est triviale. Dans notre système, les référents n'existent que par rapport aux expressions qui les dénotent et, pour une expression isolée, l'assignation de dénotation revient à dire « cette expression dénote ce qu'elle dénote ». Par contre, il n'est pas trivial de dire, étant donné un ensemble d'expressions, que telle expression dénote ce que dénote telle autre expression. Dans cette optique, pour une chaîne de référence de cardinalité n, il existe $n-1$ assignations de dénotation qui ne sont pas triviales. Ce sont ces assignations de dénotation qui seront à évaluer. On remarque que ce nombre correspond au nombre minimal de « liens de coréférence » nécessaires pour définir une chaîne de référence [5]. On note également que pour une chaîne de référence de cardinalité 1 (un singleton) il n'y a pas d'assignation de dénotation non triviale $(1-1=0)$; autrement dit, il n'y a rien à évaluer [6].

De manière générale, donc, nous considérons que les expressions d'une chaîne de coréférence peuvent – au moins partiellement – être organisées hiérarchiquement selon la spécificité de leur description par rapport à leur référent. En règle générale, les expressions les plus spécifiques d'une chaîne de coréférence sont celles qui permettront l'identification du référent de la chaîne. Pour rester au plus proche de ce que serait l'interprétation par un observateur des chaînes de référence, nous exigerons donc que l'évaluation d'une analyse des reprises avec identité de dénotation prenne en compte le degré de spécificité des descriptions à l'intérieur des chaînes pour établir la correspondance entre les référents de deux annotations.

3.4 Calcul de la correspondance

La méthode que nous utilisons pour obtenir la correspondance entre les référents de deux analyses des chaînes de référence d'un texte est divisée en deux étapes : dans un premier temps, nous recherchons la correspondance optimale (dans un sens défini ci-dessous, section 3.4.2) entre les chaînes de référence d'une annotation et celles de l'autre annotation. La correspondance que nous recherchons est celle qui, nous le rappelons, nous permettra de considérer que deux chaînes de référence peuvent bien être interprétées comme dénotant le même

[5] Voir ci-dessous la présentation de la méthode d'évaluation de Vilain et al. [91], section 3.7.2 page 116.

[6] Il n'y a rien à évaluer parce qu'on ne cherche pas à évaluer la reconnaissance de l'ensemble des expressions dénotantes, mais seulement l'identification des expressions coréférentes.

référent, à partir de quoi nous pourrons évaluer les assignations de dénotation proposées dans la réponse par rapport à celles de l'annotation clé [7].

3.4.1 Similarité entre deux chaînes de références

Pour établir une correspondance optimale entre les chaînes de référence de deux annotations, nous devons définir une mesure de la similarité entre ces chaînes. En accord avec les observations que nous avons faites dans la section précédente sur l'interprétation des chaînes de référence obtenues dans la clé et la réponse, une telle mesure de similarité doit être basée, au moins partiellement, sur les expressions que les chaînes de référence ont en commun [8]. Cependant, comme toutes les expressions d'une chaîne de référence ne contribuent pas de la même manière à l'identification de son référent, nous faisons usage d'une hiérarchie qui prend en compte la spécificité du contenu descriptif des expressions : chaque chaîne de référence S dans l'une ou l'autre des annotations est partitionnée en sous-ensembles, tels que chaque sous-ensemble contient des expressions jugées avoir un certain degré de spécificité. Nous nous baserons pour la présente description de notre mesure de similarité sur la configuration adoptée pour le test d'opérationalité qui sera présenté au chapitre suivant, à savoir une partition des chaînes de référence en cinq sous-ensembles, soit du plus spécifique au moins spécifique :

- $NP(S)$, ensemble des noms propres de S,
- $SN(S)$, ensemble des syntagmes nominaux descriptifs de S (c'est-à-dire ayant un noyau nominal ou adjectival),
- $PH(S)$, ensemble des propositions, phrases ou groupes de phrases de S,
- $PRO(S)$, ensemble des pronoms de S, sauf réfléchis et relatifs,
- $R(S)$, ensemble des pronoms réfléchis et relatifs de S.

On peut envisager plus ou moins d'ensembles selon la hiérarchie qu'on voudra se donner ; on peut aussi adopter une hiérarchie différente. Il importe cependant de noter qu'il faut se donner des critères qui soient les plus opérationnels possibles ; c'est pourquoi, par exemple, nous ne faisons pas de distinction entre différents types de syntagmes nominaux descriptifs, même si certains auront un contenu descriptif qui les rapprochera plutôt des noms propres, d'autres plutôt des pronoms. Une telle distinction, dans la mesure où elle risquerait d'être sujette à controverse, serait susceptible d'introduire un flou dans les critères d'évaluation.

[7] La modélisation du problème en termes de similarité entre chaînes de référence et correspondance optimale, ainsi que les diverses méthodes et formules permettant d'aboutir au résultat souhaité, telles qu'elles sont présentées dans la présente section, sont dues à Éric Gaussier.

[8] Étant donné deux chaînes A_{o_i} et $B_{o_{i'}}$ dans deux annotations distinctes d'un même texte, si A_{o_i} et $B_{o_{i'}}$ n'ont aucune expression en commun, il n'y a pas lieu d'envisager qu'elles puissent dénoter le même référent.

La hiérarchie que nous avons adoptée vise donc à traiter le cas général. Si une chaîne de référence contient à la fois un nom propre et un syntagme nominal descriptif, nous considérons que le nom propre est le plus spécifique. Les syntagmes nominaux descriptifs sont placés au-dessus des propositions ou phrases dans la hiérarchie. Les deux types d'expressions peuvent cependant être considérés comme équivalents. Plus qu'une notion de spécificité ici, ce qui a guidé notre choix est le fait que, selon nos définitions, plusieurs syntagmes nominaux descriptifs peuvent apparaître dans une chaîne de référence alors qu'une chaîne de référence ne peut contenir qu'une seule phrase. En d'autres termes, si on a dans l'annotation clé la chaîne de référence suivante :

- $C_{o1} = \{PH_1, SN_1, SN_2, SN_3\}$

et dans la réponse les deux chaînes :

- $R_{o1'} = \{PH_1\}$
- $R_{o2'} = \{SN_1, SN_2, SN_3\}$

nous considérerons qu'il y a une similarité plus grande entre C_{o1} et $R_{o2'}$ (ce qui ne serait pas le cas si PH_1 était un nom propre). On place enfin en bas de l'échelle de spécificité les pronoms en distinguant parmi les pronoms les réfléchis et relatifs comme étant les moins spécifiques. L'idée ici est que les réfléchis et relatifs sont plus liés au contexte textuel, en particulier par la syntaxe, que les autres pronoms. On pourra éventuellement trouver des chaînes de coréférence qui ne contiennent que des pronoms ; il nous semble impossible de trouver une chaîne de coréférence qui ne contienne que des réfléchis ou relatifs.

Étant donné deux chaînes de référence A_{o_i} et B_{o_j}, nous exigeons de notre mesure de similarité qu'elle satisfasse les conditions suivantes :

i. Si A_{o_i} et B_{o_j} sont identiques, alors leur similarité est égale à 1. Si elles n'ont aucun élément en commun, alors leur similarité est égale à 0.
ii. La similarité entre A_{o_i} et B_{o_j} doit d'abord être basée sur la similarité entre $NP(A_{o_i})$ et $NP(B_{o_j})$, puis sur la similarité entre $SN(A_{o_i})$ et $SN(B_{o_j})$, puis sur la similarité entre $PH(A_{o_i})$ et $PH(B_{o_j})$, et ainsi de suite pour les autres catégories d'expressions de notre échelle de spécificité.

La première condition est une condition courante pour les mesures de similarité ; il s'agit de définir les bornes supérieures et inférieures de la mesure. La seconde condition exprime notre désir de prendre en compte le degré de spécificité des expressions par rapport à leurs référents.

La mesure que nous avons définie pour calculer la similarité entre deux chaînes de référence quelconques A_{o_i} et B_{o_j} est une combinaison linéaire de coefficients de Dice calculés pour $NP(A_{o_i})$ et $NP(B_{o_j})$, $SN(A_{o_i})$ et $SN(B_{o_j})$, $PH(A_{o_i})$ et $PH(B_{o_j})$, $PRO(A_{o_i})$ et $PRO(B_{o_j})$, et enfin $R(A_{o_i})$ et $R(B_{o_j})$. Cette mesure est

$$sim(A_{o_i}, B_{o_j}) = \begin{cases} \varepsilon & \text{si } A_{o_i} = \emptyset \text{ ou } B_{o_i} = \emptyset, \\ 0 & \text{si } A_{o_i} \neq \emptyset \text{ ou } B_{o_i} \neq \emptyset, \text{ et } A_{o_i} \cap B_{o_i} = \emptyset, \\ \sum_k \alpha_k \dfrac{2 \times |f_k(A_{o_i}) \cap f_k(B_{o_j})|}{|f_k(A_{o_i})| + |f_k(B_{o_j})|} & \text{sinon.} \end{cases}$$

FIG. 3.2 – Mesure de similarité entre deux chaînes de référence.

donnée par la formule présentée figure 3.2, dans laquelle :

- ε est arbitrairement petit : la correspondance d'une chaîne de référence avec l'ensemble vide est préférable à la correspondance entre deux chaînes qui n'ont rien en commun,
- $|\mathcal{A}|$ est le cardinal de l'ensemble $\mathcal{A}$,
- on a cinq fonctions $f_1(X)$, $f_2(X)$, $f_3(X)$, $f_4(X)$ et $f_5(X)$ correspondant respectivement à $NP(X)$, $SN(X)$, $PH(X)$, $PRO(X)$ et $R(X)$
- et cinq valeurs $\alpha_1, \ldots, \alpha_5$ associées chacune à une des cinq fonctions f_i.

Les valeurs α_i sont des poids qui satisfont, d'une part, la contrainte $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + \alpha_5 = 1$ et, d'autre part, la contrainte $\alpha_1 > \alpha_2 > \alpha_3 > \alpha_4 > \alpha_5$, cette dernière contrainte reflétant l'exigence posée par la condition (ii) sur la spécificité des expressions [9]. Pour notre part, nous adopterons en outre une lecture forte de cette condition en exigeant : $\alpha_1 > \alpha_2 + \alpha_3 + \alpha_4 + \alpha_5$, $\alpha_2 > \alpha_3 + \alpha_4 + \alpha_5$ et $\alpha_3 > \alpha_4 + \alpha_5$. Par ailleurs, afin de satisfaire la condition (i) posée ci-dessus, nous posons, dans le calcul de la similarité entre deux chaînes, que $\frac{0}{0} = 1$; autrement dit, si les deux chaînes dont on veut calculer la similarité ne contiennent aucune expression de catégorie C, leur similarité est maximale en ce qui concerne cette catégorie.

Pour déterminer les valeurs des poids que nous utiliserons, nous avons construit un jeu de test à partir de textes réels et d'exemples inventés, ces derniers pour tester les résultats de notre mesure dans des situations extrêmes. Nous avons ensuite choisi, arbitrairement, les valeurs suivantes : $\alpha_1 = 0,60$, $\alpha_2 = 0,21$, $\alpha_3 = 0,15$, $\alpha_4 = 0,03$ et $\alpha_5 = 0,01$, qui nous donnaient les résultats souhaités. D'autres choix sont possibles, mais nous pensons que sur des exemples réels, tout jeu de valeurs établi dans le respect des contraintes devrait conduire au même résultat en ce qui concerne la correspondance entre la clé et la réponse.

La mesure de similarité proposée ici nécessite bien entendu que nous disposions de l'information sur la catégorie des expressions. Idéalement, cette informa-

[9] L'échelle de spécificité est marquée par l'ordre des indices : α_1 est associé à $f_1(X)$ qui retourne l'ensemble des expressions les plus spécifiques de X ; à l'autre extrémité de l'échelle, α_5 est associé à $f_5(X)$ qui retourne l'ensemble des expressions les moins spécifiques de X.

tion devrait être fournie manuellement pour l'ensemble des expressions des textes sur lesquels on voudra procéder à l'évaluation. Même si des outils de traitement automatique des langues (analyseurs syntaxiques de surface, outil de reconnaissance des noms propres) pourront aider dans cette tâche, celle-ci pourra s'avérer trop coûteuse si le corpus d'évaluation est important. À titre d'alternative, la variante suivante peut être utilisée : on suppose qu'on dispose au moins de l'information sur la catégorie des expressions qui appartiennent à une chaîne de coréférence dans l'annotation de référence, et on utilise en lieu et place de $|NP(B_{o_j})|$, $|SN(B_{o_j})|$, $|PH(B_{o_j})|$, $|PRO(B_{o_j})|$ et $|R(B_{o_j})|$ dans la mesure ci-dessus, le cardinal de l'ensemble de la chaîne de référence B_{o_j}. Dans ce cas, la condition qui veut que deux chaînes identiques aient une similarité de 1 n'est plus remplie. Les mesures de similarité entre chaînes pourront varier sensiblement par rapport à celles qui sont obtenues avec la formule de départ, mais nous conjecturons que la correspondance finale n'en sera pas sensiblement modifiée.

3.4.2 Correspondance entre chaînes de référence

Une fois que les mesures de similarité ont été calculées pour toutes les paires de chaînes possibles, on recherche la correspondance optimale, c'est-à-dire la correspondance qui maximise la mesure de similarité totale pour les deux annotations. Nous recherchons donc la correspondance $\mathcal{C}$, de type 1-1, qui vérifie :

$$\max_{\mathcal{C}} \sum_{(A_{o_i}, B_{o_j}) \in \mathcal{C}} sim(A_{o_i}, B_{o_j})$$

Plusieurs algorithmes peuvent être utilisés pour trouver la correspondance maximale, ou du moins s'en approcher. Une heuristique largement utilisée consiste à classer les paires (A_{o_i}, B_{o_j}) selon l'ordre décroissant de leurs mesures de similarité, puis, en procédant par itérations :

- sélectionner la meilleure paire,
- ajouter cette paire à la correspondance,
- éliminer de la liste des paires restantes toute paire qui contient un des éléments de la paire sélectionnée.

Dans le cas où des relations dénotationnelles autres que l'identité (`membre-de`, `partie-de`, etc.) sont identifiées dans les annotations, la stratégie peut être raffinée en tenant compte, dans le cas où les mesures de similarité pour deux paires distinctes sont égales, des éventuelles relations entre référents. S'il existe dans l'annotation clé une relation entre les référents de deux chaînes A_{o_i} et A_{o_j} et si $sim(A_{o_i}, B_{o_i}) = sim(A_{o_i}, B_{o_j})$, on préférera mettre en correspondance avec A_{o_i} la chaîne pour laquelle il existe dans la réponse une relation avec B_{o_k} correspondant à A_{o_j}.

3.5 Exemple

3.5.1 Une réponse fictive

Nous supposerons, pour illustrer, d'une part, le calcul de la correspondance entre une annotation clé et une réponse, puis, d'autre part, les mesures d'évaluation présentées ci-après, que nous voulons évaluer l'annotation donnée figure 3.3 pour l'exemple donné figure 3.1 page 98.

> En apportant < son$_{(1)}$ `[o1]`> appui au principe de la lutte contre l'inflation$_{(1)}$, quatre jours après avoir engagé le débat sur le thème du dialogue social, < M. François Mitterrand `[o1]`> a donné la preuve que la position du < gouvernement$_{(1)}$ `[o2]`>, expliquée depuis la fin de la semaine dernière par < M. Chirac `[o3]`>, n'est pas facile à attaquer. Certes, < le président de la République `[o3]`> ne renie nullement < son$_{(2)}$ `[o3]`> geste du 1er janvier lorsqu'< il$_{(1)}$ `[o3]`> avait reçu des représentants des cheminots en grève, et < il$_{(2)}$ `[o3]`> maintient < sa `[o3]`> critique d'une rigueur inégalement partagée. < Il$_{(3)}$ `[o3]`> reste que, parti sur la défense de la « cohésion sociale », < il$_{(4)}$ `[o3]`> a jugé prudent de < s' `[o3]`>affirmer « en phase » avec < le gouvernement$_{(2)}$ `[o2]`>, pour une fois, sur la fermeté face au risque d'une relance de l'inflation$_{(2)}$.

FIG. 3.3 – Extrait 1. Annotation à évaluer (« réponse »).

D'après cette annotation, le texte contient trois chaînes de coréférence :

R_{o1} = { *son*$_{(1)}$, *M. François Mitterrand* }
R_{o2} = { *le gouvernement*$_{(1)}$, *le gouvernement*$_{(2)}$ }
R_{o3} = { *M. Chirac, le président de la République, son*$_{(2)}$, *il*$_{(1)}$, *il*$_{(2)}$, *sa, Il*$_{(3)}$, *il*$_{(4)}$, *s'* }

On note plusieurs différences avec l'annotation clé : les expressions *l'inflation*$_{(1)}$ et *l'inflation*$_{(2)}$ n'appartiennent pas à une chaîne de coréférence ; le pronom *Il*$_{(3)}$ fait partie d'une chaîne de coréférence, de même que l'expression *M. Chirac* ; enfin les expressions *son*$_{(1)}$ et *M. François Mitterrand* constituent à elles deux une chaîne de coréférence alors qu'elle font partie d'une chaîne contenant neuf éléments dans la clé ; les sept éléments qui auraient dus être regroupés avec ces deux expressions constituent une chaîne de coréférence avec les expressions *Il*$_{(3)}$ et *M. Chirac*.

3.5.2 Correspondance entre la clé et la réponse

Il s'agit donc de mesurer la similarité entre les chaîne de référence de la clé et celles de la réponse. Nous envisageons ici les différentes mesures de similarité

entre chaînes pour lesquelles la valeur est différente de 0. Les chaînes C_{o3} et R_{o2} sont identiques, on a donc :

$$sim(C_{o3}, R_{o2}) = 1$$

La seule chaîne de référence de la clé avec laquelle la chaîne R_{o1} ait une similarité non nulle est la chaîne C_{o1} :

$$sim(C_{o1}, R_{o1}) = 0,6 \times \frac{2}{2} + 0,21 + 0,15 + 0,03 \times \frac{2}{7} + 0,01 = 0,979$$

Une autre chaîne de la réponse, R_{o3}, a une similarité non nulle avec C_{o1} :

$$sim(C_{o1}, R_{o3}) = 0,6 \times \frac{0}{2} + 0,21 \times \frac{2}{2} + 0,15 + 0,03 \times \frac{10}{12} + 0,01 \times \frac{2}{2} = 0,395$$

La chaîne R_{o3} a aussi une similarité non nulle avec la chaîne C_{o6}, constituée de la seule expression *M. Chirac* :

$$sim(C_{o6}, R_{o3}) = 0,6 \times \frac{2}{2} + 0,21 \times \frac{0}{1} + 0,15 + 0,03 \times \frac{0}{6} + 0,01 \times \frac{0}{1} = 0,75$$

ainsi qu'avec la chaîne constituée par le seul pronom $Il_{(3)}$, mettons C_{o7} [10] :

$$sim(C_{o7}, R_{o3}) = 0,6 \times \frac{0}{1} + 0,21 \times \frac{0}{1} + 0,15 + 0,03 \times \frac{2}{7} + 0,01 \times \frac{0}{1} = 0,159$$

Enfin la chaîne C_{o2} a une similarité non nulle avec deux chaînes de référence de la réponse, celle constituée par le singleton $\{l'inflation_{(1)}\}$, mettons R_{o4}, et celle constituée par le singleton $\{l'inflation_{(2)}\}$, mettons R_{o5}. La similarité est la même pour les deux couples :

$$sim(C_{o2}, R_{o4}) = sim(C_{o2}, R_{o5}) = 0,6 + 0,21 \times \frac{2}{3} + 0,15 + 0,03 + 0,01 = 0,93$$

Dans ce cas, le choix de l'une ou l'autre correspondance sera arbitraire.

Étant donné ces mesures de similarité, la correspondance optimale est celle qui est présentée figure 3.4 page suivante.

En suivant l'heuristique présentée ci-dessus (section 3.4.2), on établit d'abord la correspondance entre C_{o3} et R_{o2}, puis la correspondance entre C_{o1} et R_{o1}. La correspondance entre C_{o1} et R_{o3} est alors exclue. On établit ensuite la correspondance entre C_{o6} et R_{o3}, puis la correspondance entre C_{o2} et R_{o4}. Enfin, on note que l'une des deux occurrences de *l'inflation* dans la réponse correspond à l'ensemble vide dans la clé et que, de la même manière, le singleton $\{Il_{(3)}\}$ dans la clé correspond à l'ensemble vide dans la réponse.

[10] Dans ce cas, il ne s'agit pas à proprement parler d'une chaîne de référence, puisque le pronom est impersonnel et ne désigne aucun référent. Nous considérons néanmoins que la chaîne de référence R_{o1} est susceptible de correspondre au singleton $\{Il_{(3)}\}$.

C_{o3} = { *le gouvernement*$_{(1)}$, *le gouvernement*$_{(2)}$ }	R_{o2} = { *le gouvernement*$_{(1)}$, *le gouvernement*$_{(2)}$ }
C_{o1} = { *M. François Mitterrand*, *son*$_{(1)}$, *le président de la République*, *son*$_{(2)}$, *il*$_{(1)}$, *il*$_{(2)}$, *sa*, *il*$_{(4)}$, *s'* }	R_{o1} = { *M. François Mitterrand*, *son*$_{(1)}$ }
C_{o6} = { *M. Chirac* }	R_{o3} = { *M. Chirac*, *le président de la République*, *son*$_{(2)}$, *il*$_{(1)}$, *il*$_{(2)}$, *sa*, *Il*$_{(3)}$, *il*$_{(4)}$, *s'* }
C_{o2} = { *l'inflation*$_{(1)}$, *l'inflation*$_{(2)}$ }	R_{o4} = {*l'inflation*$_{(1)}$ }
$\emptyset$	R_{o5} = {*l'inflation*$_{(2)}$ }
C_{o7} = { *Il*$_{(3)}$ }	$\emptyset$

FIG. 3.4 – Extrait 1. Correspondance entre clé et réponse.

3.5.3 Deux ensembles de prédicats d'observation

Pour chaque couple de chaînes de référence en correspondance, nous sélectionnons une expression qui appartient aux deux chaînes pour représenter leur référent commun. Nous sélectionnons aussi pour chaque chaîne associée à l'ensemble vide une expression représentant le référent de cette chaîne. Ces expressions sont soulignées dans la figure 3.4. Pour deux chaînes en correspondance, n'importe quelle expression commune peut être prise pour représenter leur référent ; nous avons cependant choisi comme expressions représentatives celles dont la présence dans les chaînes a le plus contribué à la mesure de similarité entre les deux chaînes. Pour ces expressions, comme nous l'avons indiqué plus haut (section 3.3.2), l'assignation de dénotation sera considérée comme triviale, dans la mesure où, pour établir la correspondance entre deux chaînes de référence, il faut interpréter au moins une expression de ces chaînes.

À partir de cette correspondance, on obtient les assignations de dénotation suivantes, en ignorant les assignations de dénotation triviales. Dans la clé :

- *le gouvernement*$_{(2)}$ dénote *le gouvernement*$_{(1)}$
- *son*$_{(1)}$ dénote *M. François Mitterrand*

- *le président de la République* dénote *M. François Mitterrand*
- *son*$_{(2)}$ dénote *M. François Mitterrand*
- *il*$_{(1)}$ dénote *M. François Mitterrand*
- *il*$_{(2)}$ dénote *M. François Mitterrand*
- *sa* dénote *M. François Mitterrand*
- *il*$_{(4)}$ dénote *M. François Mitterrand*
- *s'* dénote *M. François Mitterrand*
- *l'inflation*$_{(2)}$ dénote *l'inflation*$_{(1)}$

Dans la réponse :

- *le gouvernement*$_{(2)}$ dénote *le gouvernement*$_{(1)}$
- *son*$_{(1)}$ dénote *M. François Mitterrand*
- *le président de la République* dénote *M. Chirac*
- *son*$_{(2)}$ dénote *M. Chirac*
- *il*$_{(1)}$ dénote *M. Chirac*
- *il*$_{(2)}$ dénote *M. Chirac*
- *sa* dénote *M. Chirac*
- *Il*$_{(3)}$ dénote *M. Chirac*
- *il*$_{(4)}$ dénote *M. Chirac*
- *s'* dénote *M. Chirac*

On a donc maintenant deux ensembles de prédicats d'observation. L'évaluation consiste dès lors à comparer ces deux ensembles, plus précisément à juger le second par rapport au premier.

3.6 Mesures d'évaluation

Une fois que la correspondance entre les référents des deux annotations est établie, les assignations de dénotation de la réponse peuvent être comparées à celles de l'annotation clé. À partir de cette comparaison, nous calculons des mesures d'évaluation. Outre les mesures de rappel et précision, déjà évoquées au début du chapitre (page 97), nous utiliserons trois mesures destinées à fournir une analyse plus fine des erreurs : les mesures de « substitution », « sur-génération » et « sous-génération ». Ces trois mesures, définies à l'origine pour le *Named Entity Task* des Message Understanding Conferences (MUC) [38, p. 317-332], ont été adaptées par nous pour l'évaluation de la reconnaissance des chaînes de coréférence.

3.6.1 Rappel et précision

Le dénominateur dans la mesure du rappel (« *possible* ») est le nombre total d'assignation de dénotation non triviales dans l'annotation clé (autrement dit le nombre de prédicats d'observation dans la clé). Le dénominateur dans la mesure de la précision (« *effectif* ») est le nombre total d'assignation de dénotation non

triviales dans la réponse. Comme pour une chaîne de référence A_{o_i} le nombre d'assignations de dénotation non triviales est $|A_{o_i}| - 1$, on a :

$$\textit{possible} = \sum_{o_i} (|C_{o_i}| - 1)$$

$$\textit{effectif} = \sum_{o_j} (|R_{o_j}| - 1)$$

L'ensemble vide auquel une chaîne de référence de la clé et/ou de la réponse peut correspondre n'est pas une chaîne de référence : les sommes pour les valeurs *possible* et *effectif* ne sont basées que sur les référents associés à une chaîne de référence dans la clé et la réponse, respectivement.

Le numérateur pour le rappel et la précision est le même : il s'agit du nombre d'assignations de dénotation de la réponse qui sont jugées correctes. Une assignation de dénotation « e_i dénote o_j » dans la réponse est correcte si la clé contient l'assignation de dénotation « e_i dénote o_i » et o_i et o_j correspondent.

En notant $\mathcal{C}(C_{o_i})$ la chaîne de référence R_{o_j} qui correspond à la chaîne C_{o_i} et, réciproquement, $\mathcal{C}(R_{o_j})$ la chaîne de référence C_{o_i} qui correspond à la chaîne R_{o_j}, rappel et précision sont obtenus par les formules suivantes :

$$\textit{rappel} = \frac{\sum_{o_i} (|C_{o_i} \cap \mathcal{C}(C_{o_i})| - 1)}{\sum_{o_i} (|C_{o_i}| - 1)}$$

$$\textit{précision} = \frac{\sum_{o_j} (|R_{o_j} \cap \mathcal{C}(R_{o_j})| - 1)}{\sum_{o_j} (|R_{o_j}| - 1)}$$

Le numérateur est noté différemment dans les deux mesures, mais il est bien le même dans tous les cas. La différence tient simplement en ce qu'on regarde la somme des observations correctes à partir des chaînes de la clé dans le cas du rappel, et à partir des chaînes de la réponse dans le cas de la précision.

Dans l'exemple présenté dans la section précédente, on compte dix assignations de dénotation dans la clé et dix assignations de dénotation dans la réponse. Seulement deux assignations de la réponse sont correctes :

- *le gouvernement*$_{(2)}$ dénote <u>*le gouvernement*</u>$_{(1)}$
- *son*$_{(1)}$ dénote <u>*M. François Mitterrand*</u>

Pour cette annotation, le rappel et la précision sont donc :

$$\textit{rappel} = 2/10 = 0,2$$

$$\textit{précision} = 2/10 = 0,2$$

Ces deux valeurs reflètent l'idée que l'annotation proposée est erronée sur beaucoup de points ; en particulier, selon cette analyse, le texte semble parler surtout de Jacques Chirac, alors qu'il parle en fait de François Mitterrand.

3.6.2 Substitution, sur-génération, sous-génération

Le système d'évaluation que nous proposons permet d'évaluer plus précisément les erreurs grâce à trois mesures inspirées de celles utilisées pour le *Named Entity Task* défini dans MUC : les mesures de « substitution », « sur-génération » et « sous-génération ». Pour obtenir ces mesures, nous déterminons le nombre d'assignations de dénotation [11] jugées *incorrectes*, *superflues* et *manquantes*.

Un prédicat d'observation de la forme « e_i dénote o_i » dans la réponse est *incorrect* s'il existe dans la clé un prédicat d'observation de la forme « e_i dénote o_j » et o_j ne correspond pas à o_i. La situation est celle où l'expression e_i devait bien figurer dans une chaîne de coréférence, mais elle n'a pas été incluse dans la bonne chaîne de coréférence. Dans notre exemple, les assignations de dénotation suivantes sont incorrectes :

- *le président de la République* dénote *M. Chirac*
- $son_{(2)}$ dénote *M. Chirac*
- $il_{(1)}$ dénote *M. Chirac*
- $il_{(2)}$ dénote *M. Chirac*
- *sa* dénote *M. Chirac*
- $il_{(4)}$ dénote *M. Chirac*
- *s'* dénote *M. Chirac*

Toutes ces expressions auraient dû être incluses dans la chaîne de référence qui dénote François Mitterrand.

Un prédicat d'observation de la forme « e_i dénote o_i » dans la réponse est *superflu* s'il n'existe pas dans la clé de prédicat d'observation de la forme « e_i dénote o_j ». L'expression e_i est considérée après la mise en correspondance des chaînes comme représentative du référent d'une chaîne de la clé qui se trouve en correspondance avec l'ensemble vide. À une assignation de dénotation superflue correspond un échec à identifier un référent ou l'identification d'une expression non dénotante comme dénotante. Ce dernier cas se retrouve dans notre exemple, où le prédicat d'observation suivant est superflu :

- $Il_{(3)}$ dénote *M. Chirac*

Un prédicat d'observation de la forme « e_i dénote o_i » dans la clé est *manquant* s'il n'existe pas dans la réponse de prédicat d'observation de la forme « e_i dénote o_j ». L'expression e_i est considérée après la mise en correspondance des chaînes comme représentative du référent d'une chaîne de la réponse qui se trouve en correspondance avec l'ensemble vide. À une assignation de dénotation manquante correspond l'identification dans la réponse d'un référent qui n'existe pas dans la clé. Dans notre exemple, le prédicat d'observation suivant, qui se trouve dans la

[11] Il va sans dire – et il en sera ainsi jusqu'à la fin du chapitre – qu'il s'agit toujours d'assignations de dénotation *non triviales*.

clé, est manquant dans la réponse :

– *l'inflation*$_{(2)}$ dénote $\underline{\textit{l'inflation}}_{(1)}$

Du fait que cette observation n'est pas faite dans la réponse, celle-ci dit qu'il existe deux référents distincts, dénotés chacun par une occurrence de *l'inflation*. C'est un référent de trop.

La somme des assignations de dénotation jugées incorrectes, superflues ou manquantes constitue le nombre total d'erreurs (E) :

$$E = \textit{incorrect} + \textit{superflu} + \textit{manquant}$$

Le nombre total d'erreurs est le dénominateur pour les trois mesures d'évaluation des erreurs. Ces mesures se distinguent par leur numérateur, qui est le nombre d'assignations de dénotation incorrectes, superflues ou manquantes pour, respectivement, la substitution, la sur-génération et la sous-génération. Pour notre exemple, qui contient neuf erreurs, nous obtenons les valeurs suivantes :

$$\textit{substitution} = \textit{incorrect}/E = 7/9 = 0,78$$

$$\textit{sur-génération} = \textit{superflu}/E = 1/9 = 0,11$$

$$\textit{sous-génération} = \textit{manquant}/E = 1/9 = 0,11$$

De manière globale, ces trois mesures visent à évaluer la capacité qu'un système ou un observateur a d'identifier ou non les expressions qui doivent figurer dans une chaîne de coréférence, indépendamment de la question de savoir dans quelle chaîne elles doivent être placées. Une forte sur-génération indiquera une tendance à inclure dans des chaînes de coréférence des expressions qui ne devraient pas y figurer. Inversement, une forte sous-génération indiquera une tendance à ne pas inclure dans des chaînes de coréférence des expressions qui devraient y figurer. Une forte substitution indiquera que les expressions qui devaient être incluses dans des chaînes de coréférence ont bien été identifiées comme telles, mais qu'elles n'ont pas été incluses dans les bonnes chaînes de référence. C'est le cas de notre exemple.

Pour terminer notre présentation des mesures d'évaluation, nous signalons les égalités suivantes :

$$\textit{possible} = \textit{correct} + \textit{incorrect} + \textit{manquant}$$

$$\textit{effectif} = \textit{correct} + \textit{incorrect} + \textit{superflu}$$

Des observations incorrectes dans une réponse affectent à la fois la mesure du rappel et de la précision. Des observations manquantes n'affectent que le rappel. Des observations superflues n'affectent que la précision.

3.7 Mise en perspective

Après avoir présenté dans le détail notre méthode d'évaluation pour la coréférence, nous la comparons dans cette section avec trois méthodes existantes : le système développé par Vilain et al. [91] pour le *Coreference Task* de MUC-6 [44], la méthode « classes noyaux exclusifs » de Popescu-Belis [74, 71, 72] et enfin « l'algorithme B-3 » de Bagga et Baldwin [5] [12]. Nous nous concentrerons ici sur la sémantique de ces méthodes d'évaluation, plus que sur leur aspect mathématique.

3.7.1 Exemples de situations d'évaluation

Pour mieux comprendre les différentes méthodes d'évaluation, il sera utile de voir quels scores elles produisent pour différents exemples fictifs : nous utiliserons à cet effet l'exemple déjà présenté (figure 3.1 page 98 pour la clé, figure 3.3 page 108 pour la réponse), ainsi que quatre scénarios différents établis à partir de l'exemple publié dans les actes de MUC-6 [38] [13].

L'exemple publié dans les actes de MUC-6 est un article du *Wall Street Journal* pour lequel une annotation clé des chaînes de coréférence est fournie [14]. L'annotation clé contient 15 chaînes de coréférence contenant un total de 147 expressions. 50 de ces 147 expressions sont des expressions pronominales ; elles se répartissent sur 5 des 15 chaînes de coréférence. Sont imaginées les quatre situations suivantes :

1. chacune des 147 expressions qui appartiennent à une chaîne de coréférence dans la clé constitue à elle seule une chaîne de référence singleton : aucune résolution des coréférences n'est effectuée ;

2. les 147 expressions sont regroupées dans une chaîne de coréférence unique dans la réponse ;

3. les 97 expressions non pronominales sont correctement regroupées dans 15 chaînes de coréférence qui correspondent aux 15 chaînes de la clé, mais les 50 expressions pronominales sont regroupées dans une seizième chaîne de coréférence ;

4. les 97 expressions non pronominales sont correctement regroupées dans 15 chaînes de coréférence qui correspondent aux 15 chaînes de la clé, mais le système ou l'observateur n'essaie pas d'interpréter les expressions pro-

[12] Signalons que R. Passonneau [67] a proposé une méthode, basée sur celle de Vilain et al., pour évaluer l'accord entre deux annotateurs sur l'identification des chaînes de coréférence.

[13] Trois de ces quatre scénarios ont à l'origine été proposés par Popescu-Belis [74].

[14] Le fait que l'article soit en anglais ne change rien au problème.

nominales, si bien que chacune d'elles est laissée comme une chaîne de référence singleton dans la réponse.

Les mesures de rappel et de précision obtenues par chacune des méthodes considérées dans chacune des situations sont données dans le tableau 3.1 page 116. Les chiffres de la première colonne font référence aux différentes situations (0 pour notre exemple, 1 à 4 pour les quatre situations imaginées sur l'article de MUC). Les quatre colonnes suivantes donnent le rappel et la précision (chiffres gauche et droit, respectivement) pour chacune des méthodes [15]. La dernière colonne donne les mesures de sous-génération (−g), sur-génération(+g) et substitution (sub), telles qu'obtenues selon notre méthode. Le symbole « – » indique que le dénominateur pour la précision est 0 [16].

sit.	MUC		CNE		B-3		AD		−g	+g	sub
0	0,8	0,8	0,67	0,87			0,2	0,2	0,11	0,11	0,78
1	0	–	0,10	1	0,10	1	0	–	1	0	0
2	1	0,90	0,31	0,31	1	0,19	0,27	0,24	0	0,13	0,87
3	0,96	0,97	0,69	0,84	0,63	0,78	0,62	0,63	0,02	0	0,98
4	0,62	1	0,66	1	0,49	1	0,62	1	1	0	0

TAB. 3.1 – Rappel et précision selon différentes méthodes et dans différentes situations

3.7.2 L'approche par liens de Vilain et al.

Le système d'évaluation développé par Vilain et al. [91] pour la campagne d'évaluation MUC-6 est basé sur l'idée que les chaînes de coréférence sont des classes d'équivalence définies par la relation « est coréférent avec » ou « a la même dénotation que », relation que Vilain et al. appelle « IDENT » suivant les conventions du schéma d'annotation utilisé dans MUC. Une relation particulière « e_i a la même dénotation que e_j » est appelée un « lien de coréférence ». On note que ce qui permet de considérer que la relation « a la même dénotation que » définit des classes d'équivalence est le fait qu'elle soit réflexive (e_i a la même dénotation que e_i), symétrique (si e_i a la même dénotation que e_j, e_j a la même dénotation que e_i) et transitive (si e_i a la même dénotation que e_j et e_j a la même dénotation que e_k, alors e_i a la même dénotation que e_k).

Vilain et al. remarquent que pour définir une classe d'équivalence A_{o_i} de cardinalité n, il suffit d'identifier $n-1$ liens de coréférence entre les éléments de A_{o_i}.

[15] MUC fait référence à la méthode de Vilain et al., CNE à la méthode « classes noyaux exclusifs » de Popescu-Belis, B-3 à la méthode de Bagga et Baldwin et AD à notre méthode en termes d'assignations de dénotation.

[16] Pour l'exemple correspondant à la situation 0, nous sommes incapables de donner les valeurs obtenues avec la méthode B-3.

Les dénominateurs des mesures de rappel et précision sont les mêmes que dans notre système : le nombre de liens suffisant pour définir les classes d'équivalence (en d'autres termes les chaînes de coréférence) de la clé (valeur *possible*, dénominateur du rappel) et celles de la réponse (valeur *effectif*, dénominateur de la précision). Les mesures d'évaluation de Vilain et al. diffèrent des nôtres sur ce qui est considéré comme correct ou non.

Dans le système de Vilain et al., l'idée de base est de compter comme erreurs seulement le nombre minimal de liens qui doivent être ajoutés à la clé ou la réponse pour que les deux annotations deviennent identiques.

Supposons deux annotations A et B contenant respectivement une et deux chaînes de coréférence :

$$A_{o1} = \{1, 2, 3, 4, 5\}$$
$$B_{o1'} = \{1, 2, 3\}$$
$$B_{o2'} = \{4, 5\}$$

La chaîne de coréférence A_{o1} est définie, au minimum, par quatre liens de coréférence et les chaînes de coréférence $B_{o1'}$ et $B_{o2'}$ par deux et un lien, respectivement. Pour que les deux annotations soient identiques, il suffit d'ajouter un lien de coréférence entre l'un des éléments de $B_{o1'}$ et l'un des éléments de $B_{o2'}$: ce lien à ajouter constitue la seule erreur dans le système d'évaluation de Vilain et al. Si A est la clé, il s'agit d'une erreur de rappel ; si B est la clé, d'une erreur de précision.

Reprenons l'exemple développé plus haut (figure 3.4 page 110). Pour simplifier la notation, le tableau 3.2 dresse à nouveau la liste des chaînes de référence définies dans la clé et la réponse en remplaçant les expressions par des chiffres.

clé	*réponse*
C_{o3} = {1,2}	R_{o2} = {1,2}
C_{o1} = {3,4,5,6,7,8,9,10,11}	R_{o1} = {3,4}
C_{o2} = {12,13}	R_{o3} = {5,6,7,8,9,10,11,14,15}
C_{o6} = {14}	R_{o4} = {12}
C_{o7} = {15}	R_{o5} = {13}

TAB. 3.2 – Extrait 1. Clé et réponse, notation simplifiée.

Appelons S l'ensemble des expressions qui appartiennent à une chaîne de coréférence dans la clé ou la réponse : $S = \{1, 2, ..., 15\}$. La clé et la réponse constituent des partitions de S. Pour identifier les erreurs, on se donne comme objectif, en ajoutant un nombre minimal de liens de coréférence dans la clé ou la réponse, de faire en sorte que les deux partitions soient identiques. Dans le cas présent, on obtient ce résultat en ajoutant, d'une part, deux liens de coréférence dans la clé, un lien entre 14 et un des éléments de C_{o1} et un lien entre 15 et un des

éléments de C_{o1}, et, d'autre part, deux liens de coréférence dans la réponse, un lien entre 12 et 13 et un lien entre un des éléments de R_{o1} et un des éléments de R_{o3}. Les deux partitions sont alors toutes deux : {1,2}, {3,4,5,6,7,8,9,10,11,14,15}, {12,13} [17]. Les liens rajoutés dans la clé sont des erreurs de précision (la réponse contient des liens de coréférence en trop) ; les liens rajoutés dans la réponse sont des erreurs de rappel (des liens de la clé n'apparaissent pas dans la réponse). Nous noterons le nombre d'erreurs de rappel et de précision ER et EP, respectivement.

Le chiffre correspondant aux liens de coréférence correctement identifiés (valeur *correct*) est le nombre minimal de liens nécessaires pour définir la clé moins le nombre d'erreurs de rappel :

$$correct = possible - ER$$

La mesure de rappel est alors :

$$rappel = \frac{possible - ER}{possible}$$

La mesure de précision est définie par Vilain et al. comme :

$$précision = \frac{effectif - EP}{effectif}$$

Popescu-Belis [72, p. 262] a montré que le numérateur des deux mesures était le même ; on peut donc aussi bien écrire :

$$précision = \frac{possible - ER}{effectif}$$

Dans notre exemple, la clé est définie par dix liens de coréférence. Puisqu'il n'a fallu ajouter que deux liens dans la réponse, cela signifie que huit des dix liens de la clé ont été correctement identifiés dans la réponse. On a donc un rappel et une précision de 8/10.

Si on compare notre système d'évaluation avec celui de Vilain et al., on remarque que notre système produira des scores toujours inférieurs ou égaux à ceux produits par les mesures de Vilain et al. À une assignation de dénotation correcte dans notre système correspond toujours un lien de coréférence correct dans le système de Vilain et al. Par contre, un lien de coréférence correct selon Vilain et al. n'implique pas une assignation de dénotation correcte dans notre système. Dans notre exemple, le fait, par exemple, que les expressions $il_{(1)}$ et $il_{(2)}$ appartiennent à la même chaîne de coréférence dans la réponse donne lieu à un

[17] Quand on en vient au calcul des mesures d'évaluation lui-même, le système d'évaluation de Vilain et al. n'a pas l'aspect procédural qui ressort de cette présentation. La sémantique de ce système est néanmoins celle qui est présentée ici.

lien de coréférence correct pour Vilain et al., alors que nous considérons que ces deux expressions n'ont pas reçu la bonne dénotation.

On note que les désaccords entre la mesure de Vilain et al. et la nôtre portent toujours sur ce que nous jugeons *incorrect*, et non sur ce que nous jugeons *correct*, *manquant* ou *superflu*. Il s'ensuit que la mesure de Vilain et al. donnera un score égal à la nôtre dans les cas où la mesure de substitution est nulle, comme on le voit dans les situations 1 et 4. Inversement, et comme, en règle générale, les textes contiennent souvent quelques grandes chaînes de coréférence, à une forte valeur pour la mesure de substitution correspondront des scores assez hauts selon le système de Vilain et al. Cela est visible dans les situations 2 et 3.

À notre avis, le système d'évaluation de Vilain et al. mélange deux aspects différents dans l'identification des chaînes de coréférence : l'identification des expressions qui doivent être incluses dans une chaîne de coréférence, d'une part, et l'inclusion de ces expressions dans les *bonnes* chaînes de coréférence. Cela apparaît nettement sur les résultats obtenus dans les situations 3 et 4. Dans la situation 4, où les 50 expressions pronominales appartiennent chacune à un singleton, les mesures obtenues par Vilain et al. sont les mêmes que celles qu'on obtient avec notre système. Par contre, dans le cas où les 50 expressions appartiennent à une même chaîne de coréférence (situation 3), le rappel selon Vilain et al. augmente de manière significative et la précision est peu affectée, tandis que dans notre système le rappel reste constant et la précision est fortement affectée.

L'écart maximal entre les scores obtenus par les deux méthodes apparaît dans la situation 2 (147 expressions regroupées dans une même chaîne). Pour Vilain et al., le rappel est égal à 1, puisque tous les liens de coréférence de la clé ont bien été identifiés. La précision est peu affectée, puisque qu'il n'y a en fait que 14 liens de coréférence en trop dans la réponse, ceux qui font que les 15 chaînes n'en constituent qu'une.

Dans une certaine mesure, le système de Vilain et al. est donc susceptible de donner crédit à l'observateur ou la machine qui a produit la réponse du fait que les expressions qui devaient être incluses dans une chaîne de coréférence ont bien été identifiées, indépendamment du fait de savoir si la bonne dénotation leur a été assignée. Notre système distingue ces deux aspects : la capacité qu'a l'observateur ou la machine d'identifier les expressions qui doivent appartenir à une chaîne de coréférence est évaluée au moyen des trois mesures d'analyse des erreurs.

Reprenons, au terme de cette comparaison, les termes dans lesquels Vilain et al. posent l'évaluation de la résolution des coréférences :

> Les termes du rappel (respectivement de la précision) sont trouvés en calculant le nombre minimal de liens qui doivent être ajoutés à la réponse (respectivement la clé) pour faire en sorte que [les chaînes de coréférence] soit alignées.

Selon nous, poser l'évaluation en ces termes, c'est dire « que faut-il faire pour que j'obtienne le bon résultat à partir de ce qui m'est donné ? ». Nous posons

l'évaluation en termes de « est-ce que le résultat qui m'est donné est correct ou non ? ». Il est bien possible que, dans certains cas, il y ait peu de modifications à apporter à une réponse pour obtenir une réponse parfaite, mais cela ne nous semble être qu'un problème particulier dans l'évaluation. Nous pensons que la méthode d'évaluation que nous proposons permet de distinguer l'évaluation du résultat (rappel et précision) de l'évaluation des erreurs et des directions à suivre pour améliorer le résultat.

Enfin, pour résumer les objections que nous faisons contre la méthode de Vilain et al., nous dirions que l'erreur qu'elle contient est celle qui consiste à passer de la relation « a la même dénotation que », relation qui définit une équivalence entre dénotations, à une équivalence entre les expressions elles-mêmes. L'apport de notre approche est précisément de ne pas considérer les expressions comme équivalentes.

3.7.3 La méthode « classes noyaux exclusifs » de Popescu-Belis

A. Popescu-Belis et I. Robba [74], mettant en avant que les résultats obtenus par la méthode de Vilain et al. pouvaient « aller contre l'intuition » dans certains cas [18], ont proposé « trois nouvelles méthodes pour l'évaluation de la résolution des coréférences ». Ces trois méthodes ont également été présentées par A. Popescu-Belis dans sa thèse [72] et dans un article de la revue TAL [71].

La première méthode, appelée « classes noyaux », consiste à chercher pour chaque chaîne de coréférence de la clé C_{o_i} la chaîne R_{o_i} de la réponse qui lui « correspond le mieux », à savoir celle qui a la plus grande intersection avec C_{o_i}. On compte une erreur pour chacune des expressions de C_{o_i} qui n'appartient pas à R_{o_i}. Rien n'empêche qu'une chaîne R_{o_i} puisse correspondre à plusieurs chaînes de la clé ; la correspondance est ici de type 1-n.

La seconde, appelée « classes noyaux exclusifs », ajoute à la méthode précédente la contrainte d'une correspondance de type 1-1 entre les chaînes de la clé et celles de la réponse. Popescu-Belis [72, page 268] considère cette mesure plus pertinente que la précédente sur un plan cognitif — on évite qu'une chaîne de référence de la clé puisse correspondre à plusieurs chaînes de la réponse (c'est-à-dire à plusieurs référents). Dans la mesure où, d'une part, Popescu-Belis lui-même privilégie cette méthode et où, d'autre part, elle est celle qui est la plus proche de la nôtre, c'est sur elle que nous nous concentrerons.

La troisième mesure, dite « de recouvrement quantitatif », se réduit à comparer la taille des chaînes de coréférence des deux annotations, sans tenir compte des expressions qu'elles contiennent. Popescu-Belis en signale lui-même les limites, sur lesquelles nous ne reviendront pas [19].

[18] Comme nous l'avons dit, nous avons repris ici certaines des situations d'évaluation proposées à l'origine par Popescu-Belis [74].

[19] Dans sa thèse [72], Popescu-Belis propose en outre une quatrième mesure d'évaluation « fondée sur la notion d'entropie [...] et sur les études de la transmission d'information par

Dans la mesure « classes noyaux exclusifs », Popescu-Belis, comme nous, détermine une correspondance de type 1-1 entre les chaînes de référence des deux annotations et compte comme une erreur de rappel le fait qu'une expression de C_{o_i} (chaîne de la clé) n'appartienne pas à la chaîne correspondante de la réponse, et inversement pour les erreurs de précision. Les différences entre cette méthode et la nôtre résident d'une part dans la manière de calculer la correspondance entre les chaînes des deux annotations, et d'autre part, dans la manière dont sont déterminées les observations possibles, effectives et correctes.

L'heuristique utilisée par Popescu-Belis pour déterminer la correspondance entre clé et réponse est la suivante :

- sélectionner la chaîne de la clé C_{o_i} la plus grande et déterminer comme lui correspondant la chaîne de la réponse R_{o_i} qui a la plus grande intersection avec elle,
- ajouter cette paire à la correspondance — R_{o_i} ne peut dès lors plus correspondre à une autre chaîne de la clé,
- recommencer jusqu'à épuisement des chaînes de la clé.

Cette méthode, contrairement à la nôtre, ne prend pas en compte les différents niveaux de spécificité des expressions et ne garantit pas que deux chaînes en correspondance puissent bien être interprétées comme dénotant le même référent. Dans le cas de notre exemple, avec la méthode de Popescu-Belis, on obtiendrait ainsi, au lieu de la correspondance présentée figure 3.4 page 110, une correspondance qui pour les chaînes de référence C_{o1}, C_{o6}, R_{o1} et R_{o3} serait celle qui est présentée figure 3.5 page 122. Avec une telle correspondance, ce qui est correct ou non est sensiblement modifié. Le rappel et la précision, selon notre mesure, seraient tous deux de 7/10, au lieu de 2/10 précédemment.

En ce qui concerne la méthode utilisée par Popescu-Belis pour obtenir la correspondance entre les chaînes de références de la clé et celles de la réponse, on notera par ailleurs que dans la situation 3 proposée page 115, une chaîne de référence de la clé est associée à la chaîne de la réponse qui contient les 50 pronoms du texte, plutôt qu'à la chaîne qui contient toutes les expressions non pronominales de la chaîne clé — ladite chaîne contenant plus de pronoms que d'expressions non pronominales. Cela explique la différence de rappel entre la situation 3 et 4 (voir le tableau 3.1 page 116).

Dans la méthode classes noyaux exclusifs, la manière dont sont déterminées les valeurs *possible*, *effectif* et *correct* entrant dans les mesures de rappel et précision est également différente de la nôtre. Alors que nous déterminons le nombre d'assignations de dénotation à évaluer comme étant $n-1$ pour une chaîne de cardinalité n, Popescu-Belis considère qu'à une chaîne de cardinalité n correspondent n observations. Un peu de crédit est donc donné à ce que nous considérons comme des assignations de dénotation triviales. Cette singularité explique selon nous que,

un canal ». Nous n'avions pas connaissance de cette méthode à l'époque de notre travail sur l'évaluation et le temps nous a par la suite manqué pour en faire l'étude.

C_{o3} = { *le gouvernement*$_{(1)}$, *le gouvernement*$_{(2)}$ }	R_{o2} = { *le gouvernement*$_{(1)}$, *le gouvernement*$_{(2)}$ }
C_{o1} = { *M. François Mitterrand, son*$_{(1)}$, *le président de la République, son*$_{(2)}$, *il*$_{(1)}$, *il*$_{(2)}$, *sa, il*$_{(4)}$, *s'* }	R_{o3} = { *M. Chirac, le président de la République, son*$_{(2)}$, *il*$_{(1)}$, *il*$_{(2)}$, *sa, Il*$_{(3)}$, *il*$_{(4)}$, *s'* }
C_{o6} = { *M. Chirac* }	$\emptyset$
$\emptyset$	R_{o1} = { *M. François Mitterrand, son*$_{(1)}$ }
C_{o2} = { *l'inflation*$_{(1)}$, *l'inflation*$_{(2)}$ }	R_{o4} = {*l'inflation*$_{(1)}$ }
$\emptyset$	R_{o5} = {*l'inflation*$_{(2)}$ }
C_{o7} = { *Il*$_{(3)}$ }	$\emptyset$

FIG. 3.5 – Extrait 1. Correspondance selon Popescu-Belis.

comme le fait remarquer Popescu-Belis [74], le rappel et la précision ne peuvent être nuls avec cette mesure. Cela apparaît dans les résultats obtenus avec cette mesure dans la situation d'évaluation fictive 1, situation où aucune résolution des coréférences n'est effectuée dans la réponse et où le rappel avec la méthode classes noyaux exclusifs est néanmoins de 0,1. La légère différence de rappel entre la méthode classes noyaux exclusifs et notre méthode dans la situation 4 est également causée par cette différence de traitement des assignations de dénotation triviales (voir le tableau 3.1 page 116).

La mesure de rappel, dans la méthode classes noyaux exclusifs, est exprimée par la formule :

$$\textit{rappel-cne} = \frac{1}{|E|} \times \sum_{C_{o_i} \in K} |C_{o_i} \cap \mathcal{C}(C_{o_i})|$$

où E désigne l'ensemble des expressions qui appartiennent à une chaîne de coréférence dans la clé, K désigne l'ensemble des chaînes de coréférence de la clé [20]

[20] Popescu-Belis, dans sa thèse [72, page 273], précise que dans le cas où l'ensemble E_C

et, comme précédemment, $\mathcal{C}(C_{o_i})$ désigne la chaîne de la réponse qui correspond à la chaîne C_{o_i}.

La mesure de précision est exprimée par la formule :

$$\textit{précision-cne} = 1 - \frac{1}{E} \times \sum_{C_{o_i} \in K} |\mathcal{C}(C_{o_i}) - C_{o_i}|$$

Sont donc comptées comme des erreurs de précision pour une chaîne R_{o_i} de la réponse correspondant à C_{o_i}, toutes les expressions qui appartiennent à R_{o_i} et n'appartiennent pas à C_{o_i}.

On note que, alors que, habituellement, les mesures de rappel et précision ont un numérateur commun et des dénominateurs différents, ces deux mesures ont un dénominateur identique dans la méthode classes noyaux exclusifs et se distinguent par leurs numérateurs.

Pour notre exemple, étant donné la correspondance présentée figure 3.5 page ci-contre, les scores obtenus avec la méthode classes noyaux exclusifs sont [21] :

$$\textit{rappel-cne} = \frac{1}{15} \times (2 + 7 + 1) = 0,67$$

$$\textit{précision-cne} = 1 - \frac{1}{15} \times 2 = 0,87$$

Ces scores impliquent respectivement cinq erreurs de rappel et deux erreurs de précision. Les voici :

(a-b) deux erreurs, à la fois de rappel et de précision, pour n'avoir pas rattaché les expressions *M. François Mitterrand* et $son_{(1)}$ à la chaîne R_{o3},

(c) une erreur de rappel pour n'avoir pas rattaché l'expression $l'inflation_{(2)}$ à la chaîne R_{o4},

(d-e) et deux erreurs de rappel pour avoir rattaché les expressions *M. Chirac* et $Il_{(3)}$ à la chaîne R_{o3}.

On note que si on se place en termes de coréférence, les deux dernières erreurs (d-e) seraient plutôt habituellement considérées comme des erreurs de précision : le système voit trop de coréférences. On remarque également que des erreurs qui semblent a priori du même type, à savoir les erreurs (a-b-c), donnent lieu dans un cas à des erreurs de précision mais pas dans l'autre.

---

des expressions qui appartiennent à une chaîne de coréférence de la clé et l'ensemble E_R des expressions qui appartiennent à une chaîne de coréférence de la réponse ne sont pas identiques, il faut considérer pour l'ensemble E l'union de ces deux ensembles. On prend alors en compte autant de chaînes de référence singleton que nécessaire dans la clé ou la réponse.

[21] On rappelle que si on utilisait nos mesures d'évaluation avec la correspondance présentée figure 3.5, les scores seraient de 0,7 pour le rappel et la précision. Les erreurs seraient deux assignations de dénotations manquantes, pour $l'inflation_{(2)}$ et *M. François Mitterrand*, une assignation incorrecte, pour $son_{(1)}$ et deux assignations superflues, pour *M. Chirac* et $Il_{(3)}$.

Ces remarques, auxquelles s'ajoute la remarque sur l'identité des dénominateurs formulée plus haut, montrent que les mesures de la méthode classes noyaux exclusifs n'ont pas une sémantique très claire [22].

3.7.4 L'algorithme B-3 de Bagga et Baldwin

Remarquant que le système d'évaluation de Vilain et al. avait le défaut de « pénaliser les chiffres de la précision de manière égale pour tous les types d'erreurs », Bagga et Baldwin [5] ont proposé à leur tour une méthode d'évaluation, appelée « algorithme B-3 ».

L'idée de Bagga et Baldwin est que toutes les erreurs ne sont pas équivalentes. C'est le cas, par exemple, si on a une annotation clé qui contient trois chaînes de coréférence :

$$C_{o1} = \{1, 2, 3, 4, 5\}$$
$$C_{o2} = \{6, 7\}$$
$$C_{o3} = \{8, 9, 10, 11, 12\}$$

et deux annotations A et B qui contiennent chacune une erreur de précision (dans les termes de la méthode de Vilain et al.), la chaîne C_{o2} étant unifiée avec C_{o1} dans A et la chaîne C_{o3} étant unifiée avec C_{o1} dans B :

$$A_{o1} = \{1, 2, 3, 4, 5, 6, 7\}$$
$$A_{o2} = \{8, 9, 10, 11, 12\}$$
$$B_{o1} = \{1, 2, 3, 4, 5, 8, 9, 10, 11, 12\}$$
$$B_{o2} = \{6, 7\}$$

La position de Bagga et Baldwin est que la seconde erreur est plus dommageable puisqu'elle postule plus de liens de coréférence erronés que la première [23]. Or, dans les deux cas, avec la méthode de Vilain et al., le score est le même.

Pour pallier ce défaut, Bagga et Baldwin proposent une méthode qui, de manière générale, vise à pénaliser le regroupement de chaînes de coréférence importantes. Cette méthode met en œuvre deux idées nouvelles :

1. chaque expression reçoit un score pour le rappel et la précision,
2. le rappel global et la précision globale sont basés sur une moyenne pondérée des scores obtenus pour chaque expression et/ou chaque chaîne de coréférence.

Le premier point permet de considérer une expression dans le contexte de la chaîne

[22] Popescu-Belis lui-même dit que cette mesure est « assez peu parlante » [72, page 222], malgré le fait qu'il la considère par ailleurs plus pertinente sur un plan cognitif.

[23] L'annotation A identifie 10 liens de coréférence en trop : de 6 à 1, 2, 3, 4 et 5 et de 7 à 1, 2, 3, 4 et 5, et l'annotation B vingt-cinq (pour chacune des cinq expressions de C_{o3}, cinq liens, vers chacune des expressions de C_{o1}).

à laquelle elle appartient, et ainsi de normaliser sa contribution par rapport à la cardinalité de la chaîne. Cela a pour effet de différencier les différents types d'erreurs en termes de précision. Le second point n'est pas vraiment développé par les auteurs, qui semblent privilégier un système où sont assignés des poids égaux pour chaque expression et/ou chaîne de référence.

Il est certainement possible de faire usage de poids différenciés dans le système B-3, pour prendre en compte la spécificité descriptive des expressions. Les liens entre noms propres et pronoms, par exemple, pourraient être traités de manière particulière, permettant de souligner l'importance de certaines expressions. Cependant, la méthode B-3 reste, comme la méthode de Vilain et al., basée sur une évaluation des liens de coréférence indépendamment de la dénotation qui est assignée aux expressions et souffre donc à nos yeux des mêmes faiblesses.

Pour terminer, nous noterons que notre méthode pallie aussi le défaut relevé par Bagga et Baldwin dans la mesure de Vilain et al. Dans le cas où un lien de coréférence superflu dans la réponse regroupe deux chaînes de coréférence de la clé, le nombre d'assignations de dénotation erronées sera d'autant plus important que la chaîne pour laquelle on considérera le référent non identifié sera importante.

3.8 Conclusion

Nous avons présentés dans ce chapitre un nouveau critère et de nouvelles mesures d'évaluation pour la tâche d'identification des coréférences. Plutôt que de poser le problème en termes de liens à identifier entre les expressions, nous proposons une évaluation en termes d'assignation de dénotation. Pour ce faire, nous prenons en compte la spécificité relative des expressions à l'intérieur d'une même chaîne de coréférence, spécificité à partir de laquelle nous mesurons la similarité entre les chaînes de référence de chacune des deux analyses à comparer.

Notre méthode apporte plusieurs améliorations aux méthodes d'évaluation existantes pour la même tâche. Elle prend en compte la distinction entre les expressions et la dénotation des expressions : si la relation de coréférence est une relation d'équivalence, elle l'est seulement relativement à la dénotation des expressions et non aux expressions elles-mêmes (même si elles ont la même dénotation, *il* et *Jacques Chirac* sont deux expressions différentes). Par ailleurs, notre méthode distingue clairement deux aspects conceptuellement différents : la mise en correspondance de deux interprétations d'un même texte et les jugements portés par la suite sur l'une des deux analyses par rapport à l'autre. Enfin, notre méthode offre une meilleure distinction entre une évaluation qui se concentre sur le *résultat* à obtenir (mesures de rappel et précision) et une évaluation du *processus* mis en œuvre par le système pour obtenir ce résultat (mesures d'analyse des erreurs).

Chapitre 4

Test d'opérationalité

Comme nous l'avons indiqué au début du chapitre précédent, il importe de vérifier que le système de description qu'est notre typologie des reprises est « opérationnel », dans le sens où différents observateurs, étant donné les mêmes textes, utilisent bien notre système descriptif pour aboutir aux mêmes observations. L'enjeu est le suivant : dans la perspective du développement d'un système d'hypothèses décrivant plus avant le fonctionnement des phénomènes de reprises, système aboutissant éventuellement à la création d'un outil informatique permettant d'identifier les différents liens de reprise, il s'agit de montrer que les conditions d'évaluation du système existent et que ces conditions sont *externes* au système. Les conditions d'évaluation du système existeront parce qu'on aura montré que différents observateurs s'accordent sur les observations qui doivent être faites. Ces conditions d'évaluation seront externes au système d'hypothèses parce qu'on aura démontré leur existence indépendamment du système d'hypothèses lui-même ; autrement dit, ce n'est pas le système d'hypothèses qui définit ce qui est ou n'est pas une reprise, c'est un système descriptif défini préalablement au système d'hypothèses et qui est opérationnel.

Le présent chapitre donne les résultats d'une expérience visant à attester l'inter-subjectivité des observations sur les différentes relations décrites au chapitre 2, c'est-à-dire à la fois les phénomènes de reprise décrits dans les sections 2.1 à 2.5 et les phénomènes décrits dans la section 2.6, toutes choses que nous désignerons sous les termes génériques de « relations entre expressions » ou « liens entre expressions ». Cinq étudiants du GRIL ont notés les observations qu'ils faisaient sur trois articles de journaux, observations que nous comparons avec les observations que nous-mêmes avons faites par ailleurs sur ces textes. L'existence des conditions d'évaluation dépendra du degré d'accord entre les observations faites par les étudiants et nos propres observations.

Au-delà de son intérêt théorique évoqué plus haut, d'un point de vue plus pratique, celui de l'ingénierie linguistique, le test d'opérationalité que nous présentons ici nous permettra d'identifier ce qu'il est raisonnable de tenter de traiter

automatiquement et ce qui ne l'est pas. Il s'agit d'identifier les phénomènes vers lesquels les observations des lecteurs convergent, ceux-ci étant probablement plus importants dans le processus de compréhension des textes.

Enfin le travail présenté dans ce chapitre montre aussi que les notions linguistiques ici en jeu peuvent être difficiles à appréhender et que la sophistication des descriptions linguistiques trouve ses limites dans l'opérationalité.

Le chapitre est organisé comme suit. On présente d'abord les données de l'expérience : les textes utilisés, le nombre de liens que nous-mêmes avons observés, les prédicats d'évaluation (4.1). La section 4.2 décrit les mesures d'évaluation qui seront utilisées pour mesurer l'inter-subjectivité : moyenne, variance et opinion majoritaire. La présentation des résultats fait l'objet de la suite du chapitre, avec dans un premier temps une présentation globale (4.3), puis une analyse détaillée par type de relations (sections 4.4 à 4.8). Dans une section finale, nous discutons ces résultats et envisageons quelques pistes qui permettraient peut-être un plus grand succès dans une expérience future [1].

4.1 Données de l'expérience

4.1.1 Trois textes, un expert et cinq annotateurs

Trois articles du journal La Tribune des Fossés ont été donnés à cinq annotateurs, auxquels il a été demandé d'annoter les relations entre expressions qu'ils identifiaient dans ces textes. L'annotation devait être effectuée en utilisant le système d'annotation présenté au chapitre 2, c'est-à-dire qu'il fallait non seulement identifier des relations entre expressions, mais aussi les classer suivant les différents types définis.

Dans le même temps, nous produisions nous-mêmes notre annotation pour ces trois textes. Cette annotation sera considérée comme la version de référence au regard de laquelle les annotations proposées par les cinq annotateurs seront évaluées, dans la mesure où l'objectif était de vérifier que les phénomènes que nous avions nous-mêmes circonscrits pouvaient être observés et décrits comme nous le faisions par d'autres observateurs. Dans le même ordre d'idée, comme nous étions *a priori* la personne la mieux à même de manipuler le système de description proposé, nous nous donnerons le titre d'« expert ».

[1] Le travail présenté dans ce chapitre a fait l'objet d'une publication [88]. Deux erreurs affectent les chiffres présentés dans l'article publié ; elles sont corrigées ici. Les deux points sur lesquels les chiffres diffèrent sont les suivants : d'une part, une erreur dans le programme informatique calculant les mesures a conduit à des valeurs totalement erronées pour la variance, d'autre part, une erreur a été commise à l'époque dans le relevé des liens concernant les reprises par nom propre pour l'un des annotateurs. Cette dernière erreur n'a que de très légères répercussions sur le résultat global. Puisse la communauté accepter nos excuses pour ces imprécisions.

Les cinq annotateurs étaient tous des étudiants en linguistique au GRIL : trois effectuaient leur DEA et deux étaient en thèse. Aucun ne travaillait sur un sujet en rapport direct avec celui qui nous occupe ici. Certains des étudiants étaient même de nouveaux venus dans le domaine de la linguistique.

La formation des annotateurs a été la suivante :

1. présentation des différents types de relations à observer et du système d'annotation sous forme d'un cours de deux heures,
2. premier exercice d'annotation : les cinq annotateurs et l'expert cherchent à identifier ensemble les relations apparaissant dans un texte de La Tribune,
3. deuxième exercice d'annotation : les cinq annotateurs, répartis en deux équipes, doivent annoter un texte de La Tribune,
4. correction de l'exercice, les cinq annotateurs et l'expert se réunissant à cette occasion.

Au cours de cette formation, un document présentant la typologie des relations à observer et le système d'annotation (en 29 pages) a été remis aux étudiants [2]. Ce document contenait en outre une annexe qui reprenait sous forme succincte les directives pour l'utilisation du système d'annotation (en 5 pages). Par ailleurs, les annotateurs ont reçu un second document précisant certains points qui avaient pu paraître mal compris ou peu clairs au cours de la formation (3 pages).

Par convention, lorsqu'un annotateur voulait annoter un syntagme nominal, il ne devait annoter que le syntagme nominal *noyau* (par exemple *le président* dans *le président de la République*). Pour faciliter le travail des annotateurs, les trois textes étaient fournis avec un pré-balisage de l'ensemble des syntagmes noyau au moyen de chevrons. On cherchait par là à éviter que les annotateurs ne concentrent trop leur attention sur la délimitation des expressions.

Les trois articles de La Tribune utilisés pour le test sont reproduits en annexe avec l'annotation réalisée par l'expert (annexe A.1), qui constitue l'annotation de référence à laquelle les annotations proposées par les annotateurs seront comparées. Ces dernières sont reproduites dans les annexes A.2 à A.6. Les trois textes ont été choisis au hasard dans un ensemble d'articles de La Tribune — avec cependant la contrainte de choisir un texte court, un texte long et un texte de longueur moyenne, relativement à l'ensemble des textes du corpus. Les textes étaient complètement inconnus de l'expert avant la fin de la formation des annotateurs. Dans ce qui suit, nous ferons référence aux trois textes utilisés pour le test respectivement comme le texte B (comme *BNP*), le texte G (comme *Guigou*) et le texte A (comme *Allianz*).

[2] Ce document était une version préliminaire de la typologie présentée au chapitre 2. Si ces deux documents (ancienne et nouvelle version de la typologie) diffèrent, c'est essentiellement dans leur forme et la manière de présenter les différentes notions, non sur le fonds. On notera cependant que la ligne de partage entre les phénomènes de reprise et les relations caractérisées sans recours à une identité n'était pas caractérisée comme elle l'est maintenant dans les chapitres 1 et 2.

4.1.2 Instructions aux annotateurs

Les documents remis aux étudiants spécifiaient quelques instructions que nous mentionnons ici.

Tout échange entre les annotateurs sur l'interprétation de tel ou tel point de la typologie des relations à observer était interdit.

Les annotateurs avaient pour instruction de ne pas noter ce que nous appelons des reprises avec « répétition de description », telles qu'illustrées par l'exemple suivant :

(1) Marie aime les plages de l'Atlantique ; Pierre préfère les plages de la Méditerranée.

Pour toute relation notée, une des deux expressions mises en relation devait obligatoirement être un syntagme nominal ou une expression pronominale (déterminants possessifs inclus), c'est-à-dire qu'il ne fallait pas observer, par exemple, de relations entre deux phrases.

Il ne fallait pas annoter de lien entre deux expressions si celui-ci était explicite dans le texte, un lien étant explicite s'il est exprimé par une dépendance ou un ensemble de dépendances syntaxiques. La notion de lien explicite s'entend en tenant compte éventuellement des identités de dénotation. Si on a dans un texte l'expression *l'entreprise X* et, plus loin, l'expression *son président* et que l'on interprète le déterminant possessif comme coréférent avec *l'entreprise X*, alors le lien entre *son président* et *l'entreprise X* devient explicite, puisque le possessif détermine le syntagme *son président*.

4.1.3 Relations observées par l'expert

L'annotation clé effectuée par l'expert contient 202 observations, qui se répartissent comme suit, du type de relation le plus fréquent au moins fréquent [3] :

- 129 reprises avec identité de dénotation (63,8 %)
- 26 reprises avec relation `membre-de` (12,9 %)
- 23 expressions dénotant des dates (11,4 %)

[3] À ces 202 liens observés par l'expert s'ajoutent deux liens dont nous ne tiendrons pas compte dans l'évaluation : l'un est une reprise avec identité de dénotation par un pronom réfléchi qui dépend d'une reprise de description :

> *Sa production annuelle < s'élève* `[s'élever]` *> à 20 milliards de francs et <ses encours* `[o4]` *> <* `[<s' [o4]>élever]` *> à 54 milliards.* (texte B)

et une reprise par pronom relatif non comptée parce qu'il s'est avéré qu'expert et annotateurs ont tous mal interprété ce pronom :

> *D'un point de vue stratégique, cette année devrait marquer, selon Allianz, le démarrage de sa coopération dans la gestion d'actifs avec <la Dresdner Bank* `[o10]` *>, pour <laquelle* `[o10]` *> la répartition future des compétences reste en négociation.*

Nous identifions, en seconde lecture, l'antécédent de *laquelle* comme *la gestion d'actifs*.

- 11 relations de type `rel` (5,4 %)
- 7 reprises avec identité de description (3,5 %)
- 4 reprises de type « paraphrase » (2 %)
- 1 relation de type `partie-de` (0,5 %)
- 1 reprise avec relation `distingué-de` (0,5 %)

On remarque que la fréquence des identité de dénotation est très largement supérieure aux autres types de relations. Beaucoup de relations sont peu représentées, en particulier les relations `partie-de` et `distingué-de`, les reprises de type paraphrase et les reprises avec identité de description. Il est évident qu'il sera difficile d'aboutir à un résultat significatif sur l'opérationalité de ces relations. La fréquence des reprises avec identité de dénotation, en revanche, permettra, pour ce type de relation, une analyse plus détaillée des résultats en fonction du type des expressions en jeu.

4.1.4 Organisation des données

Dans la suite du chapitre, les données seront organisées comme suit.

Liens. On appelle « lien » une observation portant sur une expression e_i, que cette observation se traduise par une mise en relation avec une autre expression ou l'association d'une description spécifique à l'expression observée (ce dernier cas s'appliquant aux seules expressions temporelles).

Reprises. Nous emploierons éventuellement le terme « reprise » pour décrire, parmi les liens, les phénomènes couverts par la notion de reprise définie dans la section 1.2 et décrits dans les sections 2.1 à 2.5 et seulement ces phénomènes.

Types de liens. Nous distinguerons les types de liens suivants :

- identité de dénotation,
- identité de description,
- reprises de type paraphrase,
- interprétation des expressions temporelles,
- liens mettant en jeu un relation référentielle.

Les liens mettant en jeu un relation référentielle sont ceux qui sont caractérisés par recours à l'une des quatre relations `membre-de`, `distingué-de`, `partie-de` ou `rel`. Certains sont donc des liens de reprise (`membre-de` et `distingué-de`), les autres non (`partie-de` et `rel`).

4.1.5 Évaluation des annotations

On donne ici les différents prédicats d'évaluation pour les liens annotés par les cinq annotateurs. Nous serons amenés par la suite à procéder à une analyse des résultats par type de liens ; les prédicats d'évaluation sont conçus dans ce sens, c'est-à-dire que lorsqu'on évaluera les annotations pour un type de lien x, on fera

abstraction des liens de type y, z etc ; qui pourront par ailleurs être présentes dans l'annotation clé ou l'annotation réponse.

Les mesures d'évaluation pour chacune des cinq annotations réponse et pour chaque type de lien seront les mêmes que celles que nous avons proposées au chapitre précédent : rappel et précision, et pour l'analyse des erreurs : substitution, sur-génération et sous-génération. L'objet de la présente section est en fait de déterminer la manière dont on obtient les différentes valeurs (*correct, incorrect*, etc.) qui entrent dans les différents rapports que constituent les mesures d'évaluation.

Dans ce qui suit, nous utiliserons des symboles de la forme e_i pour faire référence à une expression, o_i pour faire référence à un référent tel qu'associé à une chaîne de référence de l'annotation clé, et o'_i pour faire référence à un référent tel qu'associé à une chaîne de référence dans une annotation réponse.

Identité de dénotation

La méthode d'évaluation est celle qui a été présentée au chapitre précédent. On rappelle que les prédicats d'observation pour ce type de reprise sont des « assignations de dénotation » non triviales.

Un prédicat d'observation de la forme « e_i dénote o'_i » dans une annotation réponse est *correct* si le prédicat « e_i dénote o_i » existe dans l'annotation clé et o'_i et o_i ont été déterminés comme correspondant l'un à l'autre [4].

Il est *incorrect* s'il existe dans la clé un prédicat de la forme « e_i dénote o_j » et o_j correspond à un référent o'_j de la réponse, distinct de o'_i (c'est-à-dire que e_i a été à juste titre placée dans une chaîne de coréférence, mais pas pas dans la bonne chaîne).

Un prédicat d'observation de la forme « e_i dénote o'_i » dans une annotation réponse est *superflu* s'il n'existe pas dans la clé de prédicat d'observation de la forme « e_i dénote o_i » (c'est-à-dire que e_i n'appartient pas à une chaîne de coréférence dans la clé).

Un prédicat d'observation de la forme « e_i dénote o_i » dans l'annotation clé est *manquant* dans la réponse s'il n'existe pas dans la réponse de prédicat d'observation de la forme « e_i dénote o'_i » (c'est-à-dire que soit e_i n'appartient pas à une chaîne de coréférence dans la réponse, soit cette expression a été prise comme représentative d'une chaîne de coréférence de la réponse qui ne correspond à aucune chaîne de coréférence de la clé).

Relations référentielles

Les prédicats d'observation pour les liens avec relation référentielle, c'est-à-dire une des relations `membre-de`, `partie-de`, `distingué-de` et `rel`, mettent en

[4] Selon la procédure décrite au chapitre précédent, section 3.4.

relation des référents et non des expressions. La méthode de mise en correspondance des référents des annotations clé et réponse nous permet de comparer les prédicats d'observation de la réponse avec ceux de la clé.

Un prédicat d'observation de la forme « o'_i *relation* o'_j » dans une annotation réponse, où *relation* tient la place d'une des quatre relations considérées ici, est *correct* s'il existe dans l'annotation clé un prédicat d'observation « o_i *relation* o_j », tel que les référents o'_i et o'_j correspondent respectivement [5] aux référents o_i et o_j et la même relation a été utilisé dans la clé et la réponse.

Un prédicat d'observation de la forme « o'_i *relation* o'_j » dans une annotation réponse est *incorrect* s'il existe dans la clé un prédicat d'observation « o_i *relation* o_j », o'_i et o'_j correspondent respectivement à o_i et o_j, mais la relation utilisée dans la réponse est différente de celle qui est utilisée dans la clé. On remarquera que la sémantique de la mesure de substitution sera un peu différente pour ces types de liens : il s'agira d'une substitution dans le type de la relation en jeu.

Un prédicat d'observation de la forme « o'_i *relation* o'_j » dans une annotation réponse est *superflu* s'il n'existe pas dans la clé de prédicat d'observation sur les deux référents correspondants.

Un prédicat d'observation de la forme « o_i *relation* o_j » dans l'annotation clé est *manquant* dans la réponse s'il n'existe pas dans la réponse de prédicat d'observation sur les deux référents correspondants.

Expressions temporelles

Dans le cas des expressions temporelles pour lesquels une description spécifique du référent pouvait être déduite et donc devait être fournie par les annotateurs, les prédicats d'observation peuvent être vus comme ayant la forme « e_i dénote la date dénotée par la description spécifique d_i ». Par exemple dans le texte B (voir l'annexe A.1), l'expression *février* dénote la date dénotée par la description spécifique *février 1998*.

Nous dirons qu'un tel prédicat d'observation dans une annotation réponse est *correct* si la clé contient la même description d_i pour e_i et *incorrect* si la clé contient une description d_j différente de d_i pour e_i. Il est à noter que, dans le cas des expressions temporelles, on pourrait vouloir moduler la notion d'*incorrect* pour les cas où l'une des deux descriptions dénote une période de temps qui inclus la seconde. On a par exemple dans l'annotation clé la description suivante :

(2) À <l'heure `[vendredi 29 mai 1998, o0]`$_{128}$> actuelle, aucune décision n'a été prise quant à d'éventuelles cessions d'actifs...

[5] Dans le cas de la relation `rel`, l'ordre des arguments n'est pas pertinent, si bien que le prédicat en question sera aussi jugé *correct* si o'_i correspond à o_j et o'_j à o_i. Cette particularité sur l'ordre des arguments de la relation `rel` vaudra pour toute référence à cette relation. Nous ne la mentionnerons plus explicitement.

et dans l'annotation réponse 2 :

(3) À <l'heure `[mai 1998]`> actuelle, aucune décision n'a été prise quant à d'éventuelles cessions d'actifs...

Selon le critère proposé ici, cette annotation est *incorrecte* ; il est bien évident qu'on pourrait la considérer au moins comme partiellement correcte. Pour les objectifs que nous nous fixons, nous nous en tiendrons à notre critère strict, tout en gardant à l'esprit un assouplissement possible, si le type de situation illustrée par l'exemple ci-dessus s'avérait trop fréquent.

Un prédicat d'observation de la forme « e_i dénote la date dénotée par la description spécifique d_i » dans la clé est *manquant* dans la réponse si aucune description n'y est donnée pour e_i. Le même type de prédicat dans la réponse est superflu si aucune description n'est donnée pour e_i dans la clé.

Identité de description et reprise de type « paraphrase »

Les identités de description et les reprises de type « paraphrase », limitées dans notre expérience aux relations anaphoriques, sont des relations entre deux expressions : une expression anaphorique e_i et son antécédent e_j, c'est-à-dire l'expression grâce à laquelle e_i peut être interprétée. Étant donné ces deux expressions, notons l'observation d'une reprise avec identité de description ou d'une reprise de type paraphrase comme des prédicats de la forme « e_i *description* e_j » et « e_i *paraphrase* e_j », respectivement, où e_i est l'expression anaphorique et e_j son antécédent.

Un prédicat d'observation de la forme « e_i *description* e_j » ou « e_i *paraphrase* e_j » dans la réponse est *correct* si le même prédicat existe dans la clé.

Un prédicat d'observation de la forme « e_i *description* e_j » ou « e_i *paraphrase* e_j » dans la réponse est *incorrect* s'il existe dans la clé un prédicat d'observation qui relie e_i à une expression e_k distincte de e_j.

Un prédicat d'observation de la forme « e_i *description* e_j » ou « e_i *paraphrase* e_j » dans la réponse est *superflu* s'il n'existe pas dans la clé de prédicat d'observation de type *description* ou *paraphrase*, respectivement, reliant ces deux expressions

Un prédicat d'observation « e_i *description* e_j » ou « e_i *paraphrase* e_j » de la clé est *manquant* dans la réponse s'il n'existe pas dans ladite réponse de prédicat reliant e_i à une autre expression par une reprise de type *description* ou *paraphrase*, respectivement.

Faux manques et faux superflus

Comme nous l'avons indiqué au début de cette présentation des prédicats d'évaluation pour les différents types de liens, les critères proposés s'appliquent pour l'analyse des réponses pour un type de lien donné, abstraction faite des

annotations qui relèvent d'un autre type. Cela entraînera la présence de ce qu'on pourrait appeler des « faux manques » et des « faux superflus ».

Considérons l'annotation clé suivante pour le pronom *en*, une reprise avec identité de description :

(4) Les magistrats auront notamment à leur disposition <des logiciels `[logiciel]`$_X$> d'instruction assistée par ordinateur. Actuellement, seuls les juges Eva Joly et Jean-Pierre Zanotto <en `[logiciel]`$_{54}$> disposent à la galerie financière de Paris.

et l'annotation réponse suivante, où la même expression est vue comme une reprise avec coréférence :

(5) Les magistrats auront notamment à leur disposition <des logiciels `[o20]`> d'instruction assistée par ordinateur. Actuellement, seuls les juges Eva Joly et Jean-Pierre Zanotto <en `[o20]`> disposent à la galerie financière de Paris.

Lorsqu'on voudra analyser les reprises avec identité de description, les critères d'évaluation ci-dessus nous conduirons à considérer que la reprise de l'annotation clé est *manquante* dans la réponse. Inversement, lorsqu'on voudra analyser les reprises avec identité de dénotation, la reprise annotée dans la réponse sera considérée comme *superflue*. Or il est bien évident, dans un cas comme dans l'autre, que l'erreur relève d'une mauvaise détermination du *type* de la reprise plus que d'un échec à voir une reprise de la clé ou d'une tendance à voir une reprise en trop.

De tels « faux manques » et « faux superflus » ne sont pas pris en compte dans nos prédicats d'évaluation, et ne le seront donc pas non plus dans les mesures que nous utiliserons pour analyser les erreurs. Nous en tiendrons cependant compte dans la discussion des dites erreurs.

4.1.6 Documents en annexe

Pour permettre au lecteur qui désirerait entrer plus avant dans les détails du test d'opérationalité que nous présentons ici, l'ensemble des données du test est reproduit en annexe (annexe A). Cette annexe contient d'une part les différentes annotations pour les trois textes et, d'autre part, sous forme de tableaux un inventaire de tous les liens annotés, soit par l'expert, soit par les annotateurs. Les observations sont classées par type de lien et éventuellement par type d'expression.

La structure de ces tableaux est la suivante. Chaque ligne identifie un lien annoté soit par l'expert, soit par au moins un des annotateurs. Chaque lien est identifié soit par un nombre, si il a été annoté par l'expert, soit par une séquence composée de « s » (comme « superflu ») suivi d'un nombre, dans le cas contraire. Ces identifiants se retrouvent sous forme d'indices dans l'annotation clé pour les

liens annotés par l'expert et dans les annotations réponses pour les autres. La seconde colonne reproduit, éventuellement de manière abrégée, les expressions concernées. La colonne *ex* donne le type de l'expression considérée, la colonne *clé* le type attribué au lien dans la clé. Les colonnes 1 à 5 donnent les jugements portés sur les annotations réponses (*correct, manque*, etc.). Ces jugements se retrouvent sous forme d'indices associés aux expressions correspondantes dans les annotations réponses. Enfin la dernière colonne (m) donne le jugement de l'« observateur idéal » tel qu'il ressort de l'opinion majoritaire [6].

Outre une meilleure compréhension des données en jeu, l'intérêt de ces documents est de permettre au lecteur d'appréhender les difficultés que nous avons rencontrées dans la réalisation de ce test d'opérationalité, difficultés qui nous ont conduits à quelques approximations.

4.1.7 Difficultés

L'analyse des annotations produites par les cinq annotateurs s'est révélée poser de nombreuses difficultés. Les cinq annotateurs avaient à leur disposition un schéma d'annotation avec une syntaxe et une sémantique précises, ainsi qu'un certain nombre de directives à respecter ; ils n'ont pas toujours respecté la syntaxe du schéma et les directives, si bien que l'interprétation de leurs annotations pose parfois problème. Par ailleurs, dans certains cas, il est manifeste que les annotateurs n'ont pas compris la sémantique du schéma d'annotation. Le prédicat d'évaluation pour telle ou telle annotation d'un lien est donc parfois déterminé sur la base d'un jugement intuitif, autorisant quelques écarts par rapport à une interprétation littérale de l'annotation. Nous ne sommes pas en mesure de spécifier clairement les écarts autorisés, mais donnons ici quelques exemples. Signalons qu'en règle général, nos choix ont plutôt été de favoriser la sévérité des jugements [7].

Dans le texte G, littéralement, l'annotateur 1 a annoté comme coréférentes (index `o25`) un ensemble d'expressions temporelles, tout en indiquant la description *année* sur la première expression.

(6) D'ici à <2000 `[année, o25]`>, le résultat net des AGF devra être porté à 5,5 milliards de francs. Pour <1998 `[o25]`>, l'acquisition des AGF doit permettre à Allianz ...

Il semble que cette notation doive être interprétée comme signifiant que toutes les expressions ayant l'index `o25` dénotent une date, mais pas nécessairement la même. Nous n'avons tout simplement pas tenu compte de cette annotation.

[6] Sur les différents jugements, voir page 140 section 9 et l'annexe A.7. Sur l'opinion majoritaire, voir plus loin la section 4.2.3.

[7] Tous les prédicats d'évaluation sont notés dans les textes et les tableaux présentés dans l'annexe A.

L'annotateur 5 a noté les reprises de description mettant en jeu des expressions monétaires d'une manière non prévue, en utilisant à la fois la notation prévue pour les identités de dénotation et celle prévue pour les identités de description :

(7) Le nouvel ensemble, baptisé BNP Lease, affiche, sur la base de 1997, un produit net bancaire de 1,7 milliard de <francs `[franc, o7]`> et un résultat avant impôts de 700 millions <`[franc, o7]`>. Sa production annuelle s' élève à 20 milliards de <francs `[o7]`> et ses encours à 54 milliards <`[franc, o7]`>.

On considère ici la réponse correcte.

L'annotateur 3 a associé deux index différents à chacune des occurrences de l'expression *le Palais de Justice* dans le texte G. L'une d'elle est jugée coréférente avec *le pôle financier*.

(8) Guigou visite les locaux du <pôle `[o3]`> financier [...] Ils auront à leur disposition quelque 23 mètres carrés par personne, contre pratiquement moitié moins auparavant au <Palais `[o24]`> de justice. [...] Cette annexe parisienne du <Palais `[o3]`> de justice ...

Il n'y a aucun moyen de comprendre ce que l'annotateur a vraiment voulu dire, la seconde occurrence de l'expression *le Palais de justice* est considérée comme mal interprétée.

Ces quelques exemples ne sont qu'un faible échantillon des annotations qui se sont révélées difficiles à interpréter. Comme nous l'avons dit, nous ne sommes pas en mesure de spécifier un critère qui justifie systématiquement les jugements que nous avons portés sur ces annotations. Il y a donc dans notre expérience une certaine approximation dans les jugements, qui conduit naturellement à une approximation dans les résultats. Les chiffres que nous décrirons par la suite doivent donc être lus avec circonspection ; ils ne peuvent être interprétés que comme indiquant une tendance, non des valeurs absolues.

Pour finir, notons que la situation décrite ici illustre la difficulté qu'il y a à mettre en place une expérience telle que celle que nous avons menée, expérience qui porte sur un sujet, l'interprétation des textes, qu'il est difficile de traduire sur le papier. L'usage d'outils informatiques permettant aux annotateurs de mieux visualiser leurs annotations (p. ex. voir l'ensemble des expressions appartenant à une chaîne de coréférence) et de vérifier la syntaxe des annotations serait sans doute un moyen de corriger quelques défauts.

4.2 Mesures d'évaluation globales

Notre but était de vérifier que la typologie des relations présentée au chapitre 2 était opérationnelle. Nous voulions vérifier que les diverses relations décrites pouvaient être effectivement observées par d'autres observateurs et décrites de la même manière.

Idéalement, pour que nous puissions dire que la typologie est complètement opérationnelle, il faudrait que :

- les annotateurs aient observé tous et seulement les liens que nous-mêmes, expert, avons observés,
- et qu'ils aient classées ces liens de la même manière que nous.

Il est cependant plus qu'improbable qu'un annotateur humain produise une annotation parfaite. La fatigue, l'ennui que peut procurer la tâche, le fait d'avoir autre chose à faire, un échec à se concentrer sur sa tâche pour quelque raison que ce soit, sont susceptibles de conduire l'annotateur à commettre des erreurs. Nous dirons donc que notre typologie est opérationnelle si les observations que nous avons faites ont dans une large mesure également été faites par les annotateurs dans leur ensemble. Pour appréhender cette « large mesure », nous utiliserons des mesures d'évaluations qui viseront donc à considérer l'ensemble des annotateurs, plutôt que chaque annotateur tour à tour.

Ces mesures d'évaluations globales sont basées sur les mesures présentées au chapitre précédent, qui ont été définies pour évaluer une seule annotation réponse par rapport à l'annotation clé. Pour évaluer l'opérationalité de notre typologie, nous utiliserons deux ensembles de mesures d'évaluation globales : des mesures moyennes et des mesures de l'opinion majoritaire. Outre ces deux ensembles de mesures, nous calculerons la variance des différentes annotations pour évaluer la dispersion des différentes réponses autour de la moyenne.

4.2.1 Moyenne

Les plus simples des mesures globales sont obtenues par calcul des mesures moyennes à partir des cinq jeux de mesures d'évaluation obtenues pour chacune des annotations réponses. Le rappel moyen (*rappel-M*) et la précision moyenne (*précision-M*) seront donc :

$$\textit{rappel-M} = 1/n \sum_{a_i} \textit{rappel}_{a_i}$$

$$\textit{précision-M} = 1/n \sum_{a_i} \textit{précision}_{a_i}$$

où a_i réfère tour à tour à chacune des cinq annotations réponses et n est le nombre d'annotateurs pour lesquels il existe une valeur pour le rappel ou la précision. La plupart du temps, on aura $n = 5$, mais si, par exemple, pour un des cinq annotateurs la mesure de précision n'est pas pertinente en raison d'un rappel nul, n sera égal à 4 pour la mesure *précision-M*. Les mesures moyennes pour l'analyse des erreurs seront calculées de manière similaire.

Les mesures moyennes permettent d'avoir une première vue globale des annotations, mais ces mesures sont sensibles aux valeurs extrêmes, c'est-à-dire au fait que certaines annotations peuvent s'écarter sensiblement de la moyenne. Supposons que la clé contiennent trois prédicats d'observations A, B et C, et que chacun des cinq annotateurs ait correctement effectué une des observations correspondantes et ait « omis » de faire les deux autres. Le rappel moyen et la précision moyenne seraient tous deux de 1/3. Ce score ne permettrait pas de faire la différence entre une situation dans laquelle les cinq annotateurs auraient tous fait la même observation, disons A, et une situation dans laquelle par exemple deux annotateurs auraient fait l'observation A, deux autres auraient fait l'observation B et le dernier l'observation C. La différence entre ces deux situations est que, dans le premier cas, l'observation A fait clairement l'objet d'une inter-subjectivité, alors que dans le second cas, aucune des trois observations ne peut être assurément considérée comme inter-subjective. Pour pallier ce défaut et compléter notre système d'évaluation, nous calculerons, d'une part, la variance des différentes annotations par rapport à la moyenne et, d'autre part des mesures d'évaluation qui prennent en compte l'opinion majoritaire.

4.2.2 Variance

La variance permet d'évaluer l'écart des différentes annotations par rapport à la moyenne. Elle peut être utilisée pour mesurer l'inter-subjectivité des observations puisque si tous les observateurs font exactement les mêmes observations, elle est égale à 0, tandis que de fortes différences entre les différentes annotations entraîneront une forte valeur pour la variance.

Nous distinguons avec la variance la notion d'inter-subjectivité de celle d'opérationalité. Là où l'opérationalité, question centrale de notre expérience, concerne l'accord des annotateurs avec l'annotation clé, l'inter-subjectivité concernera l'accord des annotateurs les uns avec les autres. Le calcul de la variance sera basé seulement sur les différents prédicats d'évaluation pour les cinq annotations réponses, indépendamment de la sémantique de ces prédicats par rapport à l'annotation clé : si pour un lien A, les observations des cinq annotateurs ont toutes été jugées *correctes*, ou bien toutes *manquantes*, ou encore toutes *incorrectes*, alors les annotateurs s'accordent dans leur observation du lien A et la variance sera nulle pour ce lien. On effectuera par ailleurs une comparaison plus fine que celle qui serait basée sur les seuls prédicats d'évaluation, en tenant compte des types de liens différents qui pourraient être rencontrés dans différentes réponses. Par exemple, si en réponse à un lien de type `membre-de` entre deux référent o_i et o_j un annotateur propose une relation `partie-de` entre les deux référents correspondant et un autre une relation `rel`, ces deux réponses sont toutes considérées comme *incorrectes*, mais elles marquent cependant un désaccord entre les annotateurs.

La variance est définie par rapport à l'« annotation moyenne ». Les mesures

d'évaluation pour cette annotation moyenne correspondent aux mesures définies dans la section précédente, mais l'annotation moyenne est calculée pour chaque lien observé, de manière à évaluer l'écart de chacune des cinq annotations par rapport à cette moyenne lien par lien.

Considérons les liens observés dans le texte suivant, extrait du texte G [8] :

(9) Montant du loyer : <21,6 millions `[franc, o14]`> de francs par an <auxquels `[o14]`> <s' `[o15]`>ajoutent <15 millions `[franc, o15]`> de travaux spécifiques pour sécuriser les lieux <que `[o15]`> prend en charge le propriétaire de l'immeuble, la Bred.

Le tableau suivant (tableau 4.1) décrit ces liens et les prédicats d'évaluation correspondant pour chacune des cinq annotations réponses [9]. La colonne *ex* donne le type de l'expression-reprise considérée : rl = pronom relatif, rf = pronom réfléchi, sn = syntagme nominal descriptif. La colonne *clé* indique le type de lien dans l'annotation clé : ID = identité de dénotation, d = identité de description. Les colonnes *1* à *5* donnent les prédicats d'évaluation pour chacune des cinq annotations : C = correct, I = incorrect, case vide = manquant. Par exemple, le lien 42, considéré dans l'annotation clé comme une reprise de dénotation avec identité et mettant en jeu un pronom réfléchi, a été correctement observé par les annotateurs 2 et 5, n'a pas été observé par l'annotateur 3 et a été observé de manière incorrecte par les annotateurs 1 et 4.

	Liens	*ex*	*clé*	*1*	*2*	*3*	*4*	*5*
41	auxquels [s'ajoutent] → 21,6 millions	rl	ID	C	C	C	C	C
42	s'[ajoutent] → 15 millions de travaux	rf	ID	I	C		I	C
43	15 millions → 21,6 millions de F	sn	d	C			C	C
45	que [prend en charge] → 15 millions	rl	ID	I	C	C		I

TAB. 4.1 – Exemple pour le calcul de la variance

Soit e_{ai_n} le prédicat d'évaluation de l'annotation i pour le lien n. Dans le tableau ci-dessus, on a, par exemple, $e_{a1_{41}} = C$, $e_{a3_{43}} = \emptyset$ [10], $e_{a5_{45}} = I$ etc.

On définit l'annotation moyenne moy_n pour le lien n comme :

$$moy_n = (\alpha_{C_n}, \alpha_{I_n}, \alpha_{I'_n}, \alpha_{\emptyset_n}, \alpha_{ID_n}, \alpha_{ID'_n}, \alpha_{ID''_n}, \alpha_{p_n}, \alpha_{d_n}, \alpha_{M_n}, \alpha_{P_n}, \alpha_{R_n})$$

où

$$\alpha_X = \sum_{ai} \frac{1}{5}\ \delta(e_{ai_n}, X)$$

[8] Pour faciliter la lecture, seules les liens qui nous intéressent pour l'exemple sont notés.

[9] Ce tableau est extrait de ceux qui sont présentés en annexe et qui reprennent tous les liens annotés par l'expert ou les cinq annotateurs dans le test d'opérationalité. Les chiffres de la première colonne renvoient à l'annotation clé.

[10] Le symbole $\emptyset$ est utilisé pour les observations jugées *manquantes*.

avec

$$\delta(e_{ai_n}, X) = 1 \ si \ e_{ai_n} = X$$

$$\delta(e_{ai_n}, X) = 0 \ si \ e_{ai_n} \neq X$$

L'indice ai représente chacun des annotateurs tour à tour et le rapport $\frac{1}{5}$ est déterminé par le nombre d'annotateurs dans notre expérience. Chaque valeur α_{X_n} représente la proportion de prédicats d'évaluation de type X sur l'ensemble des prédicats d'évaluation pour les annotations du lien n. À titre d'exemple, pour les liens 41 et 42 l'annotation moyenne est la suivante :

$$moy_{41} = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$$

$$moy_{42} = (\frac{2}{5}, \frac{2}{5}, 0, \frac{1}{5}, 0, 0, 0, 0, 0, 0, 0, 0)$$

La variance n'est pas définie sur les seules valeurs *correct, incorrect, superflu* et *manquant* données par les prédicats d'évaluation, d'où les différentes valeurs possibles pour un prédicat d'évaluation e_{ai_n}, prédicat qui permet d'obtenir les différentes valeurs α_X dans la formule ci-dessus. On tient compte en effet non seulement des prédicats d'évaluation, mais aussi de l'accord entre les annotateurs. Supposons par exemple que pour un lien l de la clé, la réponse d'un annotateur $a1_l$ soit jugée *incorrecte* et que la réponse d'un autre annotateur $a2_l$ soit elle aussi jugée *incorrecte* ; cela n'implique nullement que les deux annotateurs aient proposé la même interprétation. Dans le cas où les deux annotateurs proposent la même interprétation, on aura $e_{a1_l} = e_{a2_l} = I$; les deux annotateurs s'accordent sur ce lien. Dans le cas où les deux interprétations diffèrent, on aura $e_{a1_l} = I$ et $e_{a2_l} = I'$; les deux annotateurs divergent et la variance sera donc plus importante. Pour une description détaillée des différentes valeurs possibles pour e_{ai_n}, on renvoie le lecteur à l'annexe A.7, qui présente sous forme de tableaux l'ensemble des liens observés par l'expert et les annotateurs avec les jugements portés sur ces derniers.

Pour un lien n, la variance est définie par la formule :

$$V_n = \frac{1}{5} \sum_{ai} (1 - m_{e_{ai_n}})^2$$

Pour le lien 42, dans notre exemple, on a les valeurs suivantes pour chacun des annotateurs :

$$m_{e_{a1_{42}}} = m_I = \frac{2}{5}$$
$$m_{e_{a2_{42}}} = m_C = \frac{2}{5}$$

$$m_{e_{a3_{42}}} = m_{\emptyset} = \frac{1}{5}$$
$$m_{e_{a4_{42}}} = m_{I} = \frac{2}{5}$$
$$m_{e_{a5_{42}}} = m_{C} = \frac{2}{5}$$

La variance pour ce lien est donc :

$$V_{42} = \frac{(1-\frac{2}{5})^2 + (1-\frac{2}{5})^2 + (1-\frac{1}{5})^2 + (1-\frac{2}{5})^2 + (1-\frac{2}{5})^2}{5} = 0,416$$

Pour les autres liens de notre exemple, on a $V_{41} = 0$, $V_{45} = 0,416$ et $V_{43} = 0,24$.

Pour calculer la variance moyenne vm_L pour un ensemble de liens L de cardinalité $|L|$, on effectue simplement la somme de toutes les variances par lien, divisée par le nombre total de liens :

$$vm_L = \frac{1}{|L|} \times \sum_{l \in L} V_l$$

Pour l'ensemble E des quatre liens de notre exemple, la variance moyenne est donc :

$$vm_E = \frac{0 + 0,416 + 0,24 + 0,416}{|E|} = \frac{1,072}{4} = 0,268$$

La variance offre une première solution en réponse aux limites des mesures d'évaluation moyennes évoquées à la fin de la section précédente. Le problème que nous évoquions est résumé dans le tableau 4.2. Dans une première situation, un lien sur trois possibles a été correctement observé par les cinq annotateurs ; dans une seconde, un lien a été correctement observé par deux annotateurs, un autre également par deux autres annotateurs et le troisième par seulement un annotateur. Dans les deux cas, tous les autres prédicats d'évaluation sont « *manquant* ».

Situation 1	*1*	*2*	*3*	*4*	*5*	*Situation 2*	*1*	*2*	*3*	*4*	*5*
lien A	C	C	C	C	C	*lien A*	C	C			
lien B						*lien B*			C	C	
lien C						*lien C*					C

TAB. 4.2 – Deux situations : mêmes rappel et précision, variances distinctes

Dans ces deux situations, le rappel moyen et la précision moyenne sont identiques, à savoir, respectivement, 0,33 et 1. Par contre, dans la première situation, la variance est nulle, alors qu'elle est égale, dans la seconde situation, à 0,214.

Pour terminer, nous signalons que pour un lien donné, la variance peut aller de 0, dans le cas où les cinq annotateurs s'accordent, à 0,64, cas où chacun des

cinq annotateurs a une interprétation singulière. Dans la mesure où le nombre d'annotateurs ayant participé au test et le nombre de prédicats d'évaluation possibles sont assez faibles, on peut s'attendre à ce que la variance soit également assez faible. Cela sera d'autant plus vrai si on considère la variance moyenne. Néanmoins, les valeurs obtenues, prises en compte de manière relative les unes par rapport aux autres, n'en seront cependant pas moins informatives et nous donneront un moyen de classer les différents types de liens selon le degré de qualité des observations faites par les annotateurs.

4.2.3 Opinion majoritaire

Avec les mesures d'évaluation moyennes, d'une part, et la variance, d'autre part, nous disposons d'une information qui devrait suffire à évaluer l'opérationalité de notre typologie des liens à observer : des valeurs fortes pour les mesures moyennes, auxquelles sera associée par nécessité une faible variance, indiqueront que la typologie est opérationnelle. Toute autre configuration des résultats mettra un doute sur l'opérationalité, mais il pourra être difficile d'analyser ce qui pose problème. Nous ferons donc usage d'une troisième mesure d'évaluation globale, qui repose sur l'opinion majoritaire qui se dégage des observations faites par les cinq annotateurs.

Prendre en compte l'opinion majoritaire revient simplement à considérer le groupe d'annotateurs comme un seul individu. Appelons cet individu « l'observateur idéal » ; c'est l'annotation de cet individu que nous évaluerons. Nous disposerons donc de mesures d'évaluation pour le groupe des annotateurs équivalentes à celles dont on dispose pour un individu particulier, c'est-à-dire qu'on pourra dire que le groupe des annotateurs s'est trompé sur tel ou tel lien, de telle ou telle manière.

Pour une observation O_i faite par l'expert, nous avons cinq prédicats d'évaluation, un pour chacune des annotations réponses. Les valeurs possibles pour ces prédicats d'évaluation ont déjà été évoquées dans la section précédente (section 4.2.2, page 141) et sont détaillées dans l'annexe A.7. Étant donné ces prédicats d'évaluation, nous obtenons l'annotation de l'observateur idéal de la manière suivante. Le cas le plus simple est celui où on a une majorité absolue, c'est-à-dire au moins trois observations jugées de la même manière : le prédicat d'évaluation qui juge ces trois observations est celui qui juge l'observation de l'observateur idéal. Dans le cas où il n'y a pas plus de deux observations qui soient jugées de manière identique, on considère la majorité relative, avec deux cas de figure :

1. si deux observations sur cinq sont jugées par le même prédicat d'évaluation e_i et les trois autres observations sont jugées chacune par un prédicat différent, le prédicat d'évaluation retenu pour l'observateur idéal est e_i,
2. si deux observations sont jugées par le même prédicat d'évaluation e_i et deux autres observations sont jugées par le même prédicat d'évaluation e_j

différent de e_i, on détermine le prédicat d'évaluation retenu pour l'observateur idéal

a. en privilégiant celui qui juge une observation effectivement exprimée par rapport à une absence d'observation (par exemple, si on a deux valeurs *correct*, deux valeurs *manquant* et une valeur *incorrect*, l'opinion majoritaire est *correcte*) ;
b. s'il y a encore égalité, en privilégiant le prédicat d'évaluation qui exprime une différence par rapport à l'annotation clé, c'est-à-dire tout prédicat d'évaluation autre que *correct* (par exemple, si on a deux valeurs *correct*, deux valeurs *incorrect* et une valeur *manquant*, l'opinion majoritaire est *incorrecte*).

En privilégiant les valeurs qui marquent un écart avec l'annotation clé, nous voulons éviter autant que possible de passer à côté des erreurs, c'est-à-dire qu'on préfère des résultats susceptibles de mettre en question l'opérationalité de notre typologie, quitte à ce que cela soit dû à la sévérité des critères d'évaluation, plutôt que des résultats qui permettraient de conclure à l'opérationalité de manière trompeuse parce que basée sur des critères trop souples. On notera que d'autres cas de figure sont *a priori* possibles, par exemple que chacune des cinq annotations soit jugée d'une manière particulière, mais, en pratique, les critères donnés ici couvrent l'ensemble des observations faites.

Reprenons, pour illustrer le mécanisme qui détermine l'opinion majoritaire, les quatre liens que nous avons utilisés pour la variance dans la section précédente. Pour le lien 41, nous avons cinq prédicats d'évaluation ayant la valeur *correct*, l'opinion majoritaire, ou annotation de l'observateur idéal, est jugée *correcte*. Pour les liens 42 et 45, on a un nombre égal de prédicats ayant la valeur *incorrect* et *correct* ; dans ce cas, on privilégie la valeur *incorrect* et l'annotation de l'observateur idéal est donc jugée *incorrecte*. Enfin, pour le lien 43, trois observations jugées *correctes* sur cinq nous donnent une annotation de l'observateur idéal *correcte*. Ces conclusions sont notées dans le tableau 4.2.3 ci-après par l'ajout d'une colonne supplémentaire — notée m pour « majorité » — à droite des colonnes dédiées aux prédicats d'évaluation sur les cinq annotations réponses.

	Liens	*ex*	*clé*	*1*	*2*	*3*	*4*	*5*	*m*
41	auxquels [s'ajoutent] → 21,6 millions	rl	ID	C	C	C	C	C	C
42	s'[ajoutent] → 15 millions de travaux	rf	ID	I	C		I	C	I
43	15 millions → 21,6 millions de F	sn	d	C			C	C	C
45	que [prend en charge] → 15 millions	rl	ID	I	C	C		I	I

TAB. 4.3 – Exemple pour le calcul de l'opinion majoritaire.

Le rappel et la précision pour l'opinion majoritaire seront calculés de la même manière que pour une annotation quelconque.

Nous constaterons dans la présentation des résultats que la mesure qui prend en compte l'opinion majoritaire a tendance à marquer plus nettement la direction prise par les mesures moyennes. Si les mesures moyennes sont relativement hautes, les mesures pour l'observateur idéal seront plus hautes, si les mesures moyennes sont basses, les mesures pour l'observateur idéal seront plus basses.

L'intérêt de l'opinion majoritaire est qu'elle permet de faire abstraction d'erreurs singulières que tel ou tel annotateur a pu commettre ici ou là — un oubli, une erreur d'interprétation que les autres annotateurs n'ont pas commise. Elle permet également de repérer précisément les endroits où les annotateurs tendent à s'écarter dans leur ensemble de l'annotation de référence et donc de repérer les points susceptibles de poser problème. Par exemple, on notera pour le lien 42 évoqué plus haut que le fait que le sujet avec lequel le relatif est coréférent soit inversé a pu conduire les annotateurs à l'erreur.

4.2.4 Seuil d'opérationalité

Nous avons dit plus haut (page 138) que nous considérerons que notre typologie est opérationnelle si les annotateurs, dans leur ensemble, et dans une large mesure, ont fait les mêmes observations que nous. Il est difficile de donner un seuil à partir duquel les mesures d'évaluation globales que nous utiliserons indiqueront que la « large mesure » est atteinte.

Dans un article [29] sur la mesure κ (Kappa) [11], devenue un standard pour évaluer l'accord entre annotateurs dans des tâches de classement, Barbara Di Eugenio a soulevé le problème de l'interprétation des valeurs obtenues avec une mesure d'évaluation en signalant que différentes échelles étaient utilisées par différentes communautés, telle communauté considérant une valeur $\kappa > 0,6$, voire $\kappa > 0,5$, comme indiquant un accord acceptable, telle autre considérant comme échec toute valeur $\kappa < 0,67$.

La difficulté d'associer une interprétation claire à une valeur donnée nous incitera à utiliser les mesures d'évaluation que nous obtiendrons de manière relative, c'est-à-dire qu'elles nous permettront de classer les différents types de liens selon qu'ils sont plus ou moins bien observés.

Enfin, puisqu'il faudra bien en venir à des conclusions sur l'opérationalité, nous considérerons, pour notre expérience, que fixer *a priori* une valeur indiquant

[11] La mesure κ vise à tenir compte du hasard dans l'accord entre plusieurs observateurs. Elle est plus spécifiquement pertinente lorsque la tâche à effectuer est une tâche de classement d'un certains nombre d'« objets » dans un certain nombre de classes. R. Passonneau [67] a proposé une méthode pour utiliser la mesure κ pour l'identification des chaînes de coréférence, mais celle-ci est basée sur la méthode d'évaluation de Vilain et al. [91] et nous nous avons pas étudié une possible application à nos propres mesures d'évaluation. Cela d'une part parce que les chiffres obtenus avec la mesure κ, s'ils sont différents, n'en seront pas plus faciles à interpréter, si l'on en croit B. Di Eugenio, d'autre part, parce que, dans la mesure où les restrictions que nous posons sont assez faibles et rendent ainsi théoriquement possibles un très grand nombre de combinaisons, la part du hasard dans le résultat final devrait être extrêmement faible.

l'opérationalité n'est pas absolument nécessaire. Dans la mesure où on dispose de mesures d'évaluation clairement spécifiées, cela peut être fait *a posteriori*, c'est-à-dire qu'au vu des résultats, nous jugerons si la proportion et/ou la nature des échecs sont susceptibles de mettre en cause ou non l'opérationalité. Il s'agira ici d'une question de bon sens ; nous ne pouvons qu'espérer que le lecteur adhérera à nos conclusions.

4.3 Vue globale des résultats

En anticipant légèrement sur les résultats qui seront présentés ci-après, nous pouvons d'ores et déjà signaler d'importantes différences dans l'observation des différents types de liens. Nous ne présenterons, par conséquent, pas d'évaluation globale pour l'ensemble de la typologie, mais plutôt successivement une évaluation pour les types de liens suivants, dans l'ordre :

- reprises de type « paraphrase »,
- reprises avec identité de description,
- reprises avec identité de dénotation,
- expressions dénotant des dates,
- liens avec relation référentielle (`membre-de`, etc.).

Pour les identités de dénotation, comme signalé plus haut, nous effectuerons une analyse plus détaillée, qui prendra en compte les différents types d'expressions (expressions pronominales, noms propres, syntagmes nominaux descriptifs, phrases et propositions).

Un ensemble de tableaux, qui recense les 202 liens observés par l'expert, les observations superflues faites par les cinq annotateurs, et donne les prédicats d'évaluation pour les observations des annotateurs qui correspondent à une observation de l'expert, est présenté en annexe (annexe A.7).

Le tableau 4.4, page 147, donne les mesures d'évaluation moyenne, d'opinion majoritaire et la variance pour chacun des cinq types de liens considérés [12]. On observe que les expressions temporelles sont les mieux observées : meilleur rappel, meilleure précision, plus faible variance. Viennent ensuite les identité de dénotation. Pour les trois autres types de liens, la F-mesure pour l'opinion majoritaire est inférieure à 0,5, ce qui ne permettra pas de conclure à l'opérationalité. Les scores sont les plus faibles pour les relations référentielles. La variance pour ces relations est cependant plus faible que pour les identités de description et les reprises de type paraphrase ; nous verrons qu'elle indique que les annotateurs se sont en quelque sorte « accordés à ne pas observer » ce type de liens.

[12] R_M, P_M et F_M notent respectivement le rappel moyen, la précision et la F-mesure moyennes; R_O, P_O et F_O ces mêmes mesures pour l'opinion majoritaire. V note la variance. Un tiret (-) dans une case indique que la valeur n'est pas calculable (aucune observation n'a été faite, on a donc une division par 0).

	R_M	P_M	F_M	R_O	P_O	F_O	V
expressions temporelles	0,76	0,86	0,81	0,87	0,95	0,91	0,162
identité de dénotation	0,70	0,81	0,75	0,84	0,92	0,88	0,181
identité de description	0,37	0,96	0,53	0,29	1	0,45	0,222
paraphrase	0,25	1	0,40	0	-	-	0,320
relations référentielles	0,14	0,25	0,18	0,13	0,71	0,22	0,184

TAB. 4.4 – Mesures d'évaluation par types de lien.

Les sections suivantes détaillent les résultats pour les différents types de liens, en commençant par les reprises avec identité de description et les reprises de type paraphrase, dans la mesure où la façon dont les annotateurs ont observé ces reprises donne lieu à des erreurs dans l'identification des coréférences. Pour chaque type de lien, un tableau donne l'ensemble des mesures d'évaluation : rappel, précision, substitution, sur-génération et sous-génération pour chaque annotateur, la moyenne et l'opinion majoritaire.

4.4 Identité de description

Le rappel moyen et la précision moyenne pour les identités de description s'élèvent respectivement à 0,37 et 0,96, avec une variance de 0,222, relativement importante par rapport aux identités de dénotation. Cette valeur s'explique par le fait que deux annotateurs ont systématiquement observé ces reprises alors que les trois autres les ont en général ignorées.

Les erreurs sont surtout des absences d'observations ou une utilisation du type identité de dénotation au lieu du type identité de description, comme l'indiquent les mesures de sous-génération pour les mesures moyennes (0,93) et l'opinion majoritaire (1).

La précision maximale est obtenue pour l'opinion majoritaire, mais avec un rappel de seulement 0,29. Sur les sept identités de description à observer, seulement deux ont été vues et correctement typées par la majorité.

	a1	a2	a3	a4	a5	moy	maj
rappel	0,71	0,14	0	0,14	0,86	0,37	0,29
précision	0,83	1	-	1	1	0,96	1
substitution	0	0	0	0	0	0	0
sur-génération	0,33	0	0	0	0	0,07	0
sous-génération	0,67	1	1	1	1	0,93	1

TAB. 4.5 – Résultats. Identité de description.

Quatre reprises ont été observées correctement par deux annotateurs (1 et 5) et « oubliées » par les trois autres. Elles mettent toutes en jeu des expressions monétaires, cas illustré par l'exemple suivant :

(10) Le nouvel ensemble, baptisé BNP Lease, affiche, sur la base de 1997, un produit net bancaire de <1,7 milliard `[franc]`$_X$> de francs et un résultat avant impôts de <700 millions `[franc]`$_{13}$>.

On voit ici une tendance à omettre ce type de reprise au fil de la lecture.

La septième identité de description figurant dans l'annotation de référence est la suivante :

(11) Les magistrats auront notamment à leur disposition <des logiciels `[logiciel]`> d'instruction assistée par ordinateur. Actuellement, seuls les juges Eva Joly et Jean-Pierre Zanotto <en `[logiciel]`$_{54}$> disposent à la galerie financière de Paris.

Les annotateurs ont bien vu la reprise mais ont tous considéré qu'il s'agissait d'une identité de dénotation. Stricto sensu, leur interprétation signifie que les logiciels dont disposent Eva Joly et Jean-Pierre Zanotto leur seront pris pour les mettre à disposition des magistrats du pôle financier. Nous ne pensons pas que cela fasse partie du projet du pôle financier, mais que chaque partie — les juges Joly et Zanotto d'un côté, les magistrats du pôle de l'autre — disposera de logiciels.

L'échec des annotateurs à observer correctement les reprises avec identité de description semble indiquer que la distinction entre identité de description et identité de dénotation n'est pas évidente à assimiler. Par ailleurs, le fait que les reprises avec identité de dénotation soient beaucoup plus fréquentes dans les textes a peut être conduit les annotateurs à négliger le cas particulier que représentent les reprises de description. Ce dernier point est une conjecture, dans la mesure où un annotateur a vu une identité de description là où le lien était d'un autre type pour l'expert.

Il faut par ailleurs noter que nous avons été amené à considérer comme correctes certaines annotations de reprises avec identité de description qui n'étaient pas réalisées suivant les directives du schéma d'annotation (voir un exemple dans la section 4.1.7). Cela ne fait que confirmer le fait que les annotateurs ont eu du mal à maîtriser ce type de reprise et la notation qui lui était associée.

4.5 Reprises de type paraphrase

Le rappel moyen pour les reprises de type paraphrase est de 0,25. Trois annotateurs n'ont jamais utilisé ce type de relation pour décrire une reprise, un n'en a vue qu'une, le dernier a correctement vu et typé les quatre reprises observées par l'expert. On notera que si les erreurs des annotateurs ont toutes été

jugées comme des observations *manquantes*, celles-ci consistent ici principalement à noter comme des coréférences les reprises de type paraphrase. Sur un total de quinze erreurs, on note deux absences d'annotation, une double erreur à la fois sur la source et le type de la reprise et douze erreurs où la reprise est vue mais considérée comme coréférence.

La variance est assez importante : 0,32. Elle reflète le comportement singulier d'une part, de l'annotateur 5, qui a bien observé ces reprises, et, d'autre part, de l'annotateur 4 qui en a « oublié » deux et s'est trompé sur l'identification de la source d'une troisième.

	a1	a2	a3	a4	a5	moy	maj
rappel	0	0,25	0	0	1	0,25	0
précision	-	1	-	-	1	1	-
substitution	0	0	0	0	-	0	0
sur-génération	0	0	0	0	-	0	0
sous-génération	1	1	1	1	-	1	1

TAB. 4.6 – Résultats. Paraphrase.

Comme pour les identités de description la faible fréquence de ce type de reprise par rapport aux identités de dénotation peut en partie expliquer le comportement des annotateurs. À cela s'ajoute également un système de notation qui pouvait favoriser les « oublis ».

Quoiqu'il en soit, la distinction entre reprises de type paraphrase et coréférence est probablement trop subtile. C'est particulièrement le cas lorsque le discours rapporté repris par le pronom est cité directement, cas où la distinction n'est plus vraiment valide, dans la mesure où les mots sont précisément ceux qui ont été prononcés, comme, par exemple, dans la phrase suivante :

(12) L'assureur allemand, qui consolide ses 51 % des AGF depuis le 1er avril, considère que <cette prise de contrôle lui confère « une très forte position dans le secteur de l'assurance mondiale, avec un pied particulièrement solide dans notre marché domestique qu'est l'Europe », `[o15]`$_X$> comme <l' `[o15(P)]`$_{92}$> explique son président, Hennig Schulte-Noelle.

Un dernier point, important, est susceptible d'expliquer l'échec des annotateurs à reconnaître les reprises de type paraphrase : la faute est en grande partie celle de l'expert — c'est-à-dire nous-mêmes. Entre le moment où nous avons rédigée la documentation destinée aux annotateurs et le test lui-même, sentant que la distinction entre reprise de type paraphrase et coréférence risquait d'être mal perçue, nous avons étendu la notion de paraphrase à toute reprise avec le pronom clitique *le*, dans laquelle le pronom peut être glosé par une proposition — par exemple dans (12), Hennig Schulte-Noelle explique *que cette prise de contrôle confère à Allianz une très forte position*, etc. Les annotateurs ont été informés

de ce changement pendant leur formation, mais la documentation n'avait pas été modifiée en conséquence. Cette négligence ne fait que confirmer la confusion qui règne sur ce type de reprise.

4.6 Identité de dénotation

4.6.1 Vue globale

Les reprises avec identité de dénotation (coréférence) sont parmi les mieux observées avec un rappel moyen de 0,70 et une précision moyenne de 0,81. La variance est relativement faible : 0,181 — un score assez proche de la valeur 0,16, qui indiquerait en moyenne un accord entre quatre annotateurs sur cinq.

En ce qui concerne les annotations particulières, on notera que les annotateurs 3 et 4 n'ont pas fait un très bon travail et s'écartent sensiblement de la moyenne. Les annotateurs 1, 2 et 5 obtiennent quant à eux des mesures d'évaluation assez proches les unes des autres et assez proches des mesures obtenues pour l'opinion majoritaire. Cette singularité des annotateurs 3 et 4 affecte les scores obtenus pour les mesures globales. À titre indicatif, les mesures moyennes calculées sur la base des annotations 1, 2 et 5 seulement seraient de 0,83 pour le rappel et 0,86 pour la précision.

	a1	a2	a3	a4	a5	moy	maj
rappel	0,78	0,82	0,42	0,61	0,88	0,70	0,84
précision	0,89	0,87	0,63	0,81	0,84	0,81	0,92
substitution	0,11	0,22	0,26	0,21	0,28	0,22	0,15
sur-génération	0,22	0,28	0,12	0,12	0,44	0,23	0,23
sous-génération	0,67	0,50	0,63	0,67	0,28	0,55	0,63

TAB. 4.7 – Résultats. Identité de dénotation.

4.6.2 Résultats par types d'expressions

La quantité de reprises avec coréférence dans le corpus utilisé pour le test nous permet d'effectuer une évaluation détaillée en fonction du type d'expression en jeu dans la reprise. Nous rappelons que notre système d'évaluation pour la coréférence est posé en termes d'« assignations de dénotation ». Une assignation de dénotation est un prédicat de la forme « e_i dénote o_i », où e_i est une expression et o_i représente un référent. Nous détaillons (tableau 4.8) les résultats selon les types d'expressions suivants pour e_i :

- expressions pronominales : pronoms réfléchis, relatifs, personnels et déterminants possessifs (64 reprises, soit 49,6 % des identités de dénotation),

– noms propres (20 reprises, soit 15,5 %),
– syntagmes nominaux descriptifs [13] (42 reprises, soit 32,6 %),
– phrases, groupes de phrases ou propositions (3 reprises, soit 2,3 %).

	R_M	P_M	F_M	R_O	P_O	F_O	V
noms propres	0,68	1	0,81	0,90	1	0,95	0,176
pronoms	0,76	0,81	0,78	0,89	0,88	0,88	0,184
sn descriptifs	0,68	0,76	0,72	0,79	0,94	0,86	0,182
phrases	0,07	1	0,13	0	-	-	0,053
total	0,70	0,81	0,75	0,84	0,92	0,88	0,164

TAB. 4.8 – Coréférence. Moyenne et opinion majoritaire par types d'expression.

Phrases

Une première remarque s'impose, au vu des résultats par types d'expression : les trois reprises avec identité de dénotation mettant en jeu une proposition, une phrase ou un groupe de phrases n'ont pas été observées. Le rappel moyen est de 0,07 ; il est nul pour l'opinion majoritaire.

Syntagmes nominaux descriptifs

Les « oublis » sont aussi ce qui affecte le rappel de l'opinion majoritaire pour les syntagmes nominaux descriptifs (0,79), comme l'indique la forte précision (0,94). Les erreurs de l'observateur idéal représenté par l'opinion majoritaire se répartissent en sept « oublis », une interprétation incorrecte, une annotation superflue et une utilisation d'une relation autre que l'identité.

L'interprétation incorrecte est la suivante : dans la phrase suivante du texte G, l'expert a considéré que *Cette annexe* faisait référence au pôle financier parisien, alors que les cinq annotateurs ont considéré dans leur majorité (4/5) que cette expression faisait référence aux *locaux* du pôle financier.

(13) <Cette annexe `[o3]`$_{62}$> parisienne du Palais de justice dédiée aux dossiers financiers devrait rapidement être suivie d'<autres pôles `[o4]`$_X$> en province.
<`[o4-dde-o3]`$_{64}$>

Nous avons considéré que les autres pôles auxquels il est fait référence dans cette phrase étaient distingués de l'être dénoté par *Cette annexe*. La description *pôle* s'appliquant alors à cet être, *Cette annexe* devait donc dénoter le pôle. Les quatre annotateurs qui ont fait un choix différent du nôtre ont bien noté que le référent

[13] Par opposition aux noms propres et expressions pronominales

de *d'autre pôles* était distingué du pôle financier, mais ils n'ont pas vu comme nous un lien entre cette expression et *Cette annexe*. Leur choix de considérer *Cette annexe* comme faisant référence aux locaux du pôle financier est probablement lié à la sémantique du nom *annexe*, plus particulièrement utilisé pour dénoter un lieu.

Ce désaccord entre les annotateurs et l'expert ne remet pas fondamentalement en cause l'opérationalité de la notion de coréférence, mais il en signale une limite dans la mesure où les deux interprétations nous semblent tout aussi valides. On peut considérer qu'il y a dans ce cas ce qu'on pourrait appeler une « proximité dénotationnelle » entre le pôle et les locaux du pôle, d'autant plus qu'il est courant de faire référence à une organisation — ici le pôle — par le lieu où elle est implantée [14].

Parmi les erreurs, on relève un cas où les annotateurs ont bien vu la reprise mais ne l'ont pas considérée comme une identité de dénotation. Ce cas concerne la coréférence qui existe selon nous entre *6 400 mètres carrés* et *quelque 23 mètres carrés par personne* dans la phrase suivante. Les deux expressions sont en effet deux manières de dénoter la même surface.

(14) <Cet immeuble `[o2]`$_{29}$> luxueux, complètement réaménagé, accueillera sur <6 400 mètres `[o11]`$_X$> carrés, d'ici à la fin de l'année, 274 magistrats et fonctionnaires, plus une trentaine d'assistants spécialisés des Finances et de la Banque de France. Ils auront à leur disposition <quelque 23 mètres `[o11]`$_{35}$> carrés par personne, contre pratiquement moitié moins auparavant au Palais de justice.

Enfin, les annotateurs se sont attachés à annoter l'expression *cette nouvelle donne* dans le texte A, alors que l'expert ne l'avait pas annotée. Nous avons considéré que si cette expression renvoyait bien au contexte, son référent n'était pas clairement identifiable comme dénoté par une expression précise. Les annotateurs ne se sont d'ailleurs pas accordé sur le référent auquel associer cette expression : un a considéré qu'il s'agissait de décentralisation, un autre de l'absence de décision quant à d'éventuelles cessions d'actifs, deux autres de la cession des 25 % de la Coface demandée par Bruxelles et le dernier n'a rien indiqué.

Noms propres

Les coréférences entre noms propres sont les mieux observées. Ce n'est pas surprenant, dans la mesure où l'identité des expressions rend les erreurs improbables. Le rappel moyen est assez faible mais cela est dû essentiellement au fait que l'annotateur 3 n'a pas annoté ces expressions (le rappel pour cet annotateur est de 0,1), suivant en cela une consigne d'annotation obsolète. Le comportement de cet annotateur explique aussi pour une bonne part une variance relativement

[14] Par exemple, *Bercy* pour le ministère des Finances, *l'Élysée* pour le président de la République et l'organisation qui l'entoure, etc.

forte pour ces expressions (0,176). La mesure de rappel de l'opinion majoritaire pour ces expressions (0,90) indique que deux erreurs ont été commises. La première consiste à ne pas voir la coréférence entre deux occurrences de *France* dans le texte A. La seconde est plus intéressante dans la mesure où elle met en lumière les limites de la notion d'identité. Étant donné :

(15) Fort du rachat des AGF, Allianz présente son nouveau visage [...] C'est un nouveau groupe Allianz qui naîtra d'ici à la fin de l'année.

l'expert a considéré que *Allianz* et *un nouveau groupe Allianz* dénotaient le même être de l'univers de dénotation, tandis que les annotateurs, dans leur majorité, ont considéré qu'il s'agissait là de deux êtres distincts — mais néanmoins reliés. Il y a clairement un lien entre les deux référents ; que ce lien soit l'identité est effectivement sujet à débat : on peut dire que la société appelée *Allianz* qui existera à la fin de l'année n'est pas la même que celle qui existe à la date de l'article, en particulier si la première est décrite comme un *nouveau* groupe, mais on peut aussi remarquer qu'elle sera le résultat d'une évolution de la seconde et, dans ce sens, qu'elles ne constituent bien qu'une seule et même entité [15].

Expressions pronominales

Les scores obtenus pour les coréférences avec pronom ou déterminant possessif, quoiqu'assez élevés, sont plus faibles que ce à quoi on pouvait s'attendre dans la mesure où ces expressions sont facilement repérables et constituent le prototype même des expressions anaphoriques. La F-mesure moyenne et la F-mesure d'opinion majoritaire pour ces expressions sont à peu près égales à la F-mesure obtenue pour l'ensemble des coréférences.

Pour détailler encore les résultats, le tableau 4.9 donne les scores obtenus pour les coréférences mettant en jeu les diverses expressions pronominales : pronoms réfléchis, pronoms relatifs, déterminants possessifs et autres pronoms.

	R_M	P_M	F_M	R_O	P_O	F_O	V
relatifs	0,52	0,71	0,60	0,58	0,78	0,67	0,281
réfléchis	0,78	0,87	0,82	0,89	0,89	0,89	0,174
possessifs	0,81	0,89	0,85	0,97	1	0,98	0,142
autres pronoms	0,81	0,70	0,74	1	0,74	0,85	0,190
total	0,76	0,81	0,78	0,89	0,88	0,88	0,184

TAB. 4.9 – Coréférence. Moyenne et opinion majoritaire pour les expressions pronominales.

[15] Cette dernière option va, selon nous, dans le sens de notre définition de l'identité de dénotation (voir la section 2.1.1).

Les déterminants possessifs sont les mieux observés. Au niveau de l'opinion majoritaire, on ne note qu'un oubli : un déterminant possessif de première personne dans le texte A (lien 188), renvoyant à « nous, Français », ensemble dont font partie l'auteur de l'article et *a priori* ses lecteurs supposés, mais auquel il n'est pas fait explicitement référence dans le texte.

En ce qui concerne les pronoms autres que réfléchis et relatifs, le rappel est parfait pour l'opinion majoritaire et la précision est de 0,74. Ce cas est le seul où la précision est inférieure au rappel. Les erreurs, au nombre de cinq, sont des erreurs de sur-génération : il s'agit des quatre reprises de type paraphrase analysées comme des coréférence par la majorité des annotateurs, et de la reprise avec identité de description évoquée plus haut (exemple (11), page 148), elle aussi annotée comme une coréférence. On notera par ailleurs que trois annotateurs sur cinq ont considéré comme une reprise le pronom *en* dans la phrase suivante, extraite du texte A.

(16) Tout en se félicitant de son acquisition, Allianz n'*en* rappelle pas moins ses objectifs.

Dans la mesure où les annotateurs ne s'accordaient pas sur la source de cette reprise, elle n'a pas été retenue comme une reprise observée par la majorité.

Sur les neuf pronoms réfléchis à observer, on relève une erreur d'interprétation au niveau de l'opinion majoritaire. Elle concerne un cas où l'inversion du sujet a sans doute induit les deux annotateurs concernés en erreur (lien 42, texte G).

Enfin, les pronoms relatifs ont été les moins bien observés, avec trois oublis et deux erreurs d'interprétation, au niveau de l'opinion majoritaire, sur douze reprises à observer. Les oublis apparaissent dans des structures qu'on peut considérer comme figées :

(17) À <ceux [o18]$_X$> <qui [o18]$_{55}$> <se [o18]$_{56}$> montrent réticents...

(18) À <l'heure [oO]$_{169}$> <où [oO]$_{170}$> il veut compter parmi les cinq leaders mondiaux de l'assurance et <où [oO]$_{173}$> il prépare son entrée...

On notera, dans le cas de ces deux occurrences du pronom *où*, qu'ils ont été pris en compte avec les identités de dénotation et non pas avec les expressions temporelles (cela en raison de l'analyse par catégorie présentée ici). Pour rester cohérent avec notre idée d'associer le bon référent aux expressions, on a cependant considéré qu'ils étaient mal interprétés si l'expression *l'heure* n'avait pas été annoté avec l'expression de date correspondant à celle de la clé.

Une des deux erreurs d'interprétation est une conséquence de l'échec à reconnaître une coréférence entre *Allianz* et *un nouveau groupe Allianz* évoqué plus haut (exemple (15) page 153). L'autre met en jeu l'exemple (19) suivant où deux annotateurs ont considéré que la source du relatif était *les lieux* plutôt que *15 millions de travaux spécifiques*, privilégiant ainsi l'antécédent le plus proche, sans doute par négligence.

(19) 21,6 millions de francs par an auxquels s'ajoutent <15 millions [o15]$_{43}$> de travaux spécifiques pour sécuriser les lieux <que [o15]$_{45}$> prend en charge le propriétaire de l'immeuble

Toutes ces erreurs ne remettent pas fondamentalement en cause l'inter-subjectivité de l'interprétation des pronoms, mais révèlent plus le manque d'attention dont ont pu faire preuve les annotateurs dans certains cas. Dans ce sens, on notera que les critères qui définissent l'opinion majoritaire (voir section 4.2.3) peuvent paraître assez sévères au regard de certaines des erreurs commises sur les expressions pronominales.

4.7 Expressions temporelles

L'annotation de référence contenait 23 expressions dénotant une date et requérant une description plus complète. Ces expressions ont été bien observées par les annotateurs, avec un rappel et une précision moyens de 0,76 et 0,86 et, pour l'opinion majoritaire, un rappel de 0,87 et une précision presque parfaite de 0,95. C'est pour ce type d'observations que la variance est la plus faible : 0,162.

Seulement deux omissions sont à noter : les adverbes *actuellement*, dans le texte G, et *hier*, dans le texte A, ont été ignorés par trois des cinq annotateurs, mais ils ont été bien observés par les deux autres.

Une erreur de précision est à relever. L'expression *ici* dans le texte suivant (lien 80) a été interprétée par nous et deux annotateurs comme dénotant la date de rédaction de l'article (29 mai 1998), mais a été interprétée par deux autres annotateurs comme dénotant le 12 juin 1998. Le cinquième annotateur ne s'étant pas prononcé, suivant les critères présentés section 4.2.3, les deux observations incorrectes ont été privilégiées par rapport aux deux observations correctes pour déterminer l'opinion majoritaire.

(20) Le développement de sa présence en France l'amènera se faire coter à Paris le 12 juin prochain. C'est un nouveau groupe Allianz qui naîtra d'ici à la fin de l'année.

	a1	a2	a3	a4	a5	moy	maj
rappel	0,74	0,65	0,78	0,70	0,91	0,76	0,87
précision	0,81	0,68	0,95	0,84	1	0,86	0,95
substitution	0,67	0,88	0,20	0,43	0	0,44	0,33
sur-génération	0	0	0	0	0	0	0
sous-génération	0,33	0,12	0,80	0,57	1	0,56	0,67

TAB. 4.10 – Résultats. Expressions temporelles.

4.8 Relations référentielles

Les liens mettant en jeu une relation référentielle (`membre-de`, `partie-de`, `distingué-de` ou `rel`) que nous avons observés dans les trois textes n'ont dans une large mesure pas été vus par les cinq annotateurs. En outre, les annotateurs, quand ils observaient ces liens, ont eu tendance à utiliser les types `membre-de`, `partie-de` et `rel` en contradiction avec l'annotation clé et les uns avec les autres.

Le rappel moyen est de 0,14 et la précision moyenne de 0,25. La variance est relativement faible : 0,184. Compte tenu des faibles scores obtenus pour les mesures moyennes, elle reflète surtout un accord *à ne pas voir* les liens de ce type. En ce qui concerne l'opinion majoritaire, le rappel et la précision s'élèvent respectivement à 0,13 et 0,71, ce qui indique que quelques liens ont bien été observés par les annotateurs.

	a1	a2	a3	a4	a5	moy	maj
rappel	0,08	0,18	0,10	0,15	0,18	0,14	0,13
précision	0,18	0,25	0,33	0,32	0,18	0,25	0,71
substitution	0,06	0,15	0,02	0,12	0,26	0,12	0,03
sur-génération	0,17	0,30	0,17	0,19	0,36	0,25	0,03
sous-génération	0,70	0,54	0,81	0,68	0,38	0,62	0,94

TAB. 4.11 – Résultats. Relations référentielles.

Sur les 39 liens mettant en jeu une relation référentielle dans l'annotation clé, seulement cinq ont été correctement observés par la majorité des annotateurs. Ces liens sont les suivants. Dans le texte G, il est fait référence d'une part à un ensemble spécifique de magistrats, d'autre part à l'ensemble des magistrats en général. Le premier est un sous-ensemble du second :

(21) <Les magistrats `[o19]`$_{49}$> auront notamment à leur disposition des logiciels d'instruction assistée par ordinateur. [...] on n'échappera pas à un besoin de spécialisation croissante des <magistrats `[o21(G)]`$_X$> en matière d'information économique et financière.
<`[o19-mde-o21]`$_{61}$>

L'observation de ce lien aura sans doute été facilitée par l'identité du noyau des deux syntagmes nominaux.

Dans ce même texte, le lien entre *moitié moins* et *quelque 23 mètres carrés par personne* a également été vu par les annotateurs dans leur ensemble, même si deux annotateurs ont proposé pour ce cas un système d'annotation original et qui donc n'était pas tout à fait correct :

(22) Ils auront à leur disposition <quelque 23 mètres `[o11]`$_{35}$> carrés par personne, contre pratiquement <moitié `[o12]`$_X$> moins auparavant au Palais de justice.
<`[o12-rel-o11]`$_{37}$>

Dans la phrase suivante, le référent de *les 25 % de la Coface* est un sous-ensemble des actifs qui doivent être cédés. L'adverbe *notamment* indique cette relation.

(23) À l'heure actuelle, aucune décision n'a été prise quant à d'éventuelles cessions d'<actifs [o37]$_X$>, et notamment <les 25 % [o38]$_X$> de la Coface, comme le demande Bruxelles.
<[o38-mde-o37]$_{129}$>

Enfin, dans le texte suivant, le référent de *le premier* est un élément de l'ensemble dénoté par *d'autres pôles*, et ce dernier est « distingué » du référent de *Cette annexe*.

(24) <Cette annexe [o3]$_{62}$> parisienne du Palais de justice dédiée aux dossiers financiers devrait rapidement être suivie d'<autres pôles [o4]$_X$> en province. <Le premier [o5]$_X$> sur la liste du gouvernement est le pôle corse.
<[o4-dde-o3]$_{64}$>
<[o5-mde-o4]$_{65}$>

L'observation de ce lien a été facilitée par le fait que la relation `distingué-de` est définie spécifiquement par référence à l'emploi de l'adjectif *autre*. Le lien est en outre confirmé par le fait que le référent de *Cette annexe* est lui-même décrit par ailleurs comme un *pôle* [16]. Enfin la reprise au niveau de *le premier* a sans doute été observée parce que l'expression a un contenu descriptif dont la faiblesse la rapproche du pronom.

Outre ces cinq liens bien observés, les annotateurs ont également observé dans leur majorité deux autres relations référentielles. Ces deux liens correspondent à des liens observés par l'expert, mais les annotateurs et l'expert sont en désaccord quant au type de lien en jeu.

Dans le texte suivant, nous avons observé une coréférence entre *6400 mètres carrés* et *quelque 23 mètres carrés par personne* — nous considérons que les deux expressions sont deux manières différentes de dénoter la surface de l'immeuble en question. Les annotateurs ont vu là une relation `partie-de`, ou bien une relation `membre-de`, ou simplement une relation non spécifiée.

(25) Cet immeuble luxueux, complètement réaménagé, accueillera sur <6 400 mètres [o11]$_X$> carrés, d'ici à la fin de l'année, 274 magistrats et fonctionnaires, plus une trentaine d'assistants spécialisés des Finances et de la Banque de France. Ils auront à leur disposition <quelque 23 mètres [o11]$_{35}$> carrés par personne, contre pratiquement moitié moins auparavant au Palais de justice.

[16] Les annotateurs ont en fait interprété *Cette annexe* comme dénotant un référent distinct du pôle financier, mais ils ont bien distingué les autres pôles du pôle financier.

Enfin, les annotateurs ont observé le lien entre les logiciels et les équipements informatiques mentionnés dans l'extrait suivant, mais seulement deux l'ont vu comme un lien de type « membre-de », deux autres le voyant comme un lien de type « partie-de » et le dernier comme une identité.

(26) Outre des effectifs regroupés et supplémentaires, le pôle financier bénéficiera d'<équipements `[o16]`$_X$> informatiques. Les magistrats auront notamment à leur disposition <des logiciels `[logiciel, o17]`$_X$> d'instruction assistée par ordinateur.
<`[o17-mde-o16]`$_{51}$>

Comme précédemment, on remarquera la présence de l'adverbe *notamment*, indicateur à nos yeux d'une reprise de type « membre-de ».

Pour terminer ces remarques sur la manière dont les annotateurs ont observés les relation référentielle, nous signalerons qu'il ne se trouve aucun cas où les annotateurs s'accordent dans leur majorité sur l'observation d'un lien que nous jugerions superflu — ce qu'indique la valeur nulle pour la sur-génération au niveau de l'opinion majoritaire. Dans la mesure où les erreurs pour l'opinion majoritaire sont surtout dues à des absences d'observation (la sous-génération est égale à 0,94), nous avons envisagé d'assouplir nos critères d'évaluation et de compter comme *corrects* les cas où deux annotateurs s'accordaient sur l'observation d'un lien, sans tenir compte de la relation référentielle choisie, tandis que les trois autres n'observaient rien. Avec ce nouveau critère, très peu exigeant, neuf observations évaluées à l'origine comme *manquantes* devenaient *correctes*, mais six observations *superflues* étaient ajoutées dans le même temps. Rien n'y fera, les liens mettant en jeu une relation référentielle ont bien été mal observés.

4.9 Discussion

Nous tirons dans cette section les conclusions de notre expérience. En ce qui concerne l'opérationalité de notre typologie, la conclusion est assez simple : seules les reprises avec coréférence et les expressions temporelles sont bien observées. Cet échec à démontrer l'opérationalité de notre typologie nous incite à envisager quelques pistes pour permettre l'obtention de meilleurs résultats dans l'avenir.

4.9.1 Notions opérationnelles

Des résultats de notre expérience nous pouvons conclure que l'identification des reprises avec identité de dénotation est inter-subjective, avec cette restriction que les reprises de type paraphrase ne soient pas distinguées, puisque nous avons vu que leur définition était mal spécifiée et qu'elles étaient mal observées (voir la section 4.5). Nous pouvons aussi conclure à l'inter-subjectivité sur l'interprétation

des expressions temporelles. L'opérationalité des autres notions constituant notre typologie n'a pas été démontrée.

Ces résultats confirment ceux qui ont été obtenus pour les expériences menées dans le cadre des campagnes d'évaluation MUC [20] [17], résultats qui ont conduit à la définition d'une tâche qui se limite à la seule identification des relations de coréférence [44]. En effet, les résultats indiquent que les conditions d'évaluation n'existent que pour la notion de coréférence et les expressions temporelles.

Ces résultats vont également dans le sens de ceux obtenus par M. Poesio et R. Vieira [70], qui ont étudié l'accord entre observateurs sur l'interprétation des syntagmes nominaux définis et ont obtenu des résultats bien meilleurs sur les coréférences que sur un type de lien correspondant à nos relations référentielles (« *bridging references* »). Nous reviendrons sur les résultats de Poesio et Vieira ci-dessous.

4.9.2 Pistes pour une amélioration des résultats

Dans une large mesure, l'objectif visé par notre expérience n'a pas été atteint : nous n'avons pu que montrer l'inter-subjectivité des observations pour les expressions temporelles et l'identité de dénotation, inter-subjectivité dont il n'y avait pas vraiment lieu de douter compte tenu des campagnes MUC. Nous essayons dans cette section de comprendre cet échec et de donner des éléments qui pourront peut-être servir à la définition de nouvelles expériences.

Notre sentiment général sur l'échec rencontré est que la tâche à accomplir par les annotateurs était trop compliquée pour être réalisée de manière fiable. Celle-ci pourrait être décomposée de plusieurs manière, de sorte que les observateurs puissent se concentrer plus précisément sur un problème particulier.

Restrictions sur les formes des expressions

Une première possibilité de décomposition est de restreindre la typologie à un ensemble spécifique d'expressions identifiées sur la base d'une information syntaxique : par exemple, les observateurs pourraient n'avoir à annoter que les syntagmes nominaux démonstratifs et définis et les pronoms, c'est-à-dire les expressions potentiellement anaphoriques (voir la section 1.1).

Une telle approche aurait l'avantage de spécifier un ensemble fini d'expressions dans un texte de telle manière qu'on pourra mener une expérience à « choix forcé », c'est-à-dire qu'on pourra demander aux observateurs une réponse pour chacune des expressions, en particulier de dire explicitement « il n'y a pas lieu de mettre cette expression en relation avec une autre ». On éviterait ainsi, dans une certaine mesure, le problème posé par ce que nous avons appelé les « oublis », c'est-à-dire une absence d'annotation dont on ne sait pas s'il faut l'interpréter

[17] L'article cité ne donne pas de résultats chiffrés mais seulement des informations générales.

comme une réponse positive signifiant qu'il n'y a pas de lien à observer au niveau de cette expression.

Notons que le fait d'envisager de restreindre l'annotation des liens aux cas où ils mettraient en jeu une expression typiquement anaphorique n'est pas à proprement parler une remise en cause de la notion de reprise en tant que généralisation de l'anaphore par abstraction sur le type d'expression en jeu (voir la définition de la notion de reprise dans la section 1.2 et la discussion de la section suivante). De meilleurs observations sur un ensemble particulier d'expression ne changerait rien au fait que la notion de reprise est indépendante de la forme des expressions.

Par ailleurs, le fait qu'on limite les observations à certaines expressions est loin de garantir de meilleurs résultats. M. Poesio et R. Vieira [70] ont mené une étude visant à évaluer, par deux expériences, l'accord entre différent observateurs sur une tâche consistant, pour la première expérience, à classer les syntagmes nominaux définis apparaissant dans un texte en 4 classes :

- *anaphoric same head* ; syntagmes définis dont le noyau est identique à celui de l'expression source, avec laquelle ils sont coréférents ;
- *associative* ; syntagmes définis interprétés par identité de dénotation mais avec noyau différent de la source ou interprétés de manière anaphorique avec une relation référentielle ;
- *larger situation/unfamiliar* ; syntagmes définis dont le référent est caractérisé relativement à la situation au sens large (p. ex. *la guerre Iran/Irak, le gouvernement*) ;
- *idioms* ; expressions idiomatiques ou emplois métaphoriques.

Les résultats obtenus par Poesio et Vieira indiquent un accord qui, s'il est nettement meilleur que celui que nous avons mesuré dans notre expérience, reste relativement faible [18]. On notera que l'accord mesuré par Poesio et Vieira est le plus faible pour la classe *associative*, qui est la plus proche de nos liens avec relation référentielle. Lorsqu'ils commentent les résultats de la seconde expérience, où une classe *bridging reference* correspondant de manière encore plus proche à nos relations référentielles est utilisée, les auteurs notent que « [les observateurs] ont été très mauvais pour distinguer les *bridging references* des autres descriptions définies. »

Restriction aux seuls liens de reprise

La notion de reprise, telle que nous l'avons définie dans la section 1.2, résulte en partie des conséquences que nous tirons de l'expérience présentée dans ce chapitre.

[18] Comme la tâche demandée aux observateurs est une tâche de classement, Poesio et Vieira utilisent la mesure κ. Les valeurs obtenues ne sont pas directement comparables aux nôtres.

Les identités de dénotation, identités de description et reprises de type « paraphrase », ont été relativement bien observées, modulo les erreurs de typage des identités de description et reprises de type « paraphrase ». Nous avons déjà évoqué les problèmes posés par ces deux types de reprises : dans les deux cas, leur faible fréquence a probablement facilité les négligences et, dans le cas des reprises de type « paraphrase », s'ajoute à cela un problème de définition.

Parmi les sept liens mettant en jeu une relation référentielle qui ont été (plus ou moins bien) observés (voir section 4.8), on recense six liens qui, selon l'expert, sont des liens de reprise : quatre relations `membre-de`, une identité de dénotation et une relation `distingué-de`. Le septième lien avec relation référentielle vue par la majorité des annotateurs est le suivant :

(22) Ils auront à leur disposition <quelque 23 mètres `[o11]`$_{35}$> carrés par personne, contre pratiquement <moitié `[o12]`$_X$> moins auparavant au Palais de justice.
<`[o12-rel-o11]`$_{37}$>

Ce lien relève de ce que Halliday et Hasan appellent la « référence comparative » (voir la section 2.7.2). Nous avons dit que, de manière générale, notre notion de reprise ne rend pas compte de ce type de lien, si ce n'est par l'identité de description, dans la mesure où on peut conjecturer que les liens de référence comparative mettent en fait en jeu une comparaison entre des êtres de même type. C'est le cas dans l'exemple (22), où les deux êtres dénotés sont des surfaces, ce que l'annotateur 4 a d'ailleurs noté :

(22) Ils auront à leur disposition <quelque 23 mètres [surface]> carrés par personne, contre pratiquement <moitié `[surface]`$_C$> moins auparavant au Palais de justice.

On pourrait donc analyser cet exemple comme mettant en jeu une identité de description implicite, sous-jacente à la référence comparative. Nous souhaiterions dans l'avenir explorer la notion de référence comparative pour poursuivre la caractérisation des liens caractérisés par recours à une relation d'identité que sont les reprises. À cet égard, nous avons défini les reprises comme des liens caractérisés en ayant recours à une *relation* d'identité, ce qui est mettre en jeu une sorte de comparaison (voir la section 1.2.1).

Nous pensons qu'une tâche restreinte à l'identification des reprises, voire aux seules identités de description, relations `membre-de` et relation `distingué-de`, puisque nous savons que les identités de dénotation sont bien observées, donnerait de meilleurs résultats que l'expérience que nous avons décrite. Deux facteurs essentiels seraient susceptibles d'influer sur cette éventuelle amélioration : d'une part, la réduction du type de liens à observer devrait permettre une meilleure concentration sur les objectifs, d'autre part, nous pensons avoir fourni dans les chapitres 1 et 2, avec la notion de reprise, une meilleure spécification des phénomènes à observer. Cela est particulièrement vrai pour la relation `membre-de` : en ce

qui concerne cette relation, bon nombre des annotations proposées indiquent que les annotateurs n'ont pas tenu compte de la nécessité d'une identité de description dans les reprises avec relation `membre-de`. Il faut noter que dans le document remis aux étudiants, cette nécessité était beaucoup moins explicitée que dans la présente thèse.

Un travail à accomplir dans le futur sera de conduire une nouvelle expérience pour attester l'opérationalité de la notion de reprise. Cette notion, selon nous, mérite qu'une telle expérience soit menée.

Généralisation du principe des descriptions complètes

Pour finir, nous voudrions proposer une manière différente de voir les relations qui ne relèvent pas de notre notion de reprise (c'est-à-dire les relations décrites dans la section 2.6.1), approche qui serait peut-être susceptible de rencontrer plus de succès que l'approche en termes de liens utilisée dans l'expérience décrite ici. La perspective que nous voulons proposer s'appuie sur les idées suivantes.

On remarque que les expressions dénotant des dates ont été bien interprétées par les annotateurs. On ne peut évaluer si ce résultat est dû à la particularité des objets désignés, ou au système de notation particulier à ces expressions (qui ne demandait pas une mise en relation, mais une description complète de la date désignée), mais notre proposition serait d'appliquer ce même type de notation pour les expressions qui ne sont pas interprétées en ayant recours à une relation d'identité.

La relation entre deux êtres de l'univers de dénotation dénotés par deux expressions indépendantes est très souvent traduisible en une relation de dépendance syntaxique. De fait, nous avons employé cette méthode pour mettre à jour certains liens dans nos exemples.

Dans la phrase suivante, le câble d'alimentation est le câble *de l'imprimante*.

(27) Ne placez pas <l'imprimante `[o1]`> à un endroit où des personnes pourraient marcher sur <le câble `[o2]`> d'alimentation.
<`[o2-pde-o1]`>

Dans le texte suivant, la cathédrale est la cathédrale *de Clermont-Fd*.

(28) <Jean `[o1]`> est allé à <Clermont-Fd `[o2]`>. <Là `[o2]`>, <il `[o1]`> a visité <la cathédrale `[o3]`>.
<`[o3-pde-o2]`>

Dans le texte suivant, les négociations sont des négociations *entre la Chase Manhattan Bank et Morgan Stanley Dean Witter sur le rachat des activités de conservation* etc.

(29) La Chase Manhattan Bank a confirmé <le rachat `[o2]`> des activités de conservation de titres et de compensation de Morgan Stanley Dean Witter. Lors des <négociations `[o1]`>, la presse avait rapporté que <la

transaction [o2]> <s' [o2]> élevait autour de 600 millions de dollars.
<[o1-pde-o2]>

Dans le texte suivant, les éléments en possession de la CGT sont des éléments *sur le Crédit Lyonnais* et les possibles conséquences catastrophiques sont des conséquences *de l'accord*.

(30) À peine confirmé, <l'accord [o1]> de principe trouvé entre Paris et Bruxelles pour <le Crédit Lyonnais [o2]> suscite déjà des remous. Le secrétaire général de la CGT, Louis Viannet, a estimé vendredi que le « diktat » du commissaire européen à la concurrence l'avait « emporté ». Dans une lettre au Premier ministre Lionel Jospin, le leader syndical met en garde contre <de possibles conséquences [o3]> « catastrophiques » pour l'emploi.
<[o3-rel-o1]>
« <Les éléments [o4]> en notre possession font en effet apparaître un total de cession d'<actifs [o5]> en Europe, dans le monde et en France d'environ 800 milliards », poursuit-il.
<[o4-rel-o2]>
<[o5-rel-o2]>

Les compléments en italiques dans les textes présentant chacun des cinq exemples ci-dessus explicitent la description des êtres dénotés par les expressions correspondantes dans les exemples, de la même manière que les descriptions complètes des dates caractérisaient précisément les référents des expressions temporelles. Plutôt que de mettre en relation les expressions du texte entre elles, une possibilité serait de spécifier des compléments tels que ceux que nous avons proposés. On aurait par exemple :

(29) La Chase Manhattan Bank a confirmé <le rachat [o2]> des activités de conservation de titres et de compensation de Morgan Stanley Dean Witter. Lors des <négociations [entre la Chase Manhattan Bank et Morgan Stanley Dean Witter sur le rachat...]>, la presse avait rapporté que <la transaction [o2]> <s' [o2]> élevait autour de 600 millions de dollars.

On notera que dans chacun des compléments en italiques, figure une expression qui est coréférente avec une expression du texte et que celle-ci est celle avec laquelle le lien entre l'expression anaphorique et sa source s'établit. Dans le cas de (29), on a même trois expressions coréférentes dans le complément, qui pourrait être réécrit ainsi :

(29) <La Chase Manhattan Bank [o3]> a confirmé <le rachat [o2]> des activités de conservation de titres et de compensation de <Morgan Stanley Dean Witter [o4]>. Lors des <négociations [entre o3 et o4 sur o2]>, la presse avait rapporté que <la transaction [o2]> <s' [o2]> élevait autour de 600 millions de dollars.

En d'autres termes, une autre approche serait de demander, pour chaque expression dénotante e_i qui n'est pas interprétée par reprise, de spécifier, dans la mesure du possible, une description complète de l'être désigné par cette expression, description telle qu'elle a pour base e_i et met en jeu dans un complément au moins une expression coréférente avec une expression du texte dans lequel figure e_i.

Ce traitement des relations référentielles permettrait de ne pas avoir à spécifier, du moins dans un premier temps, des types de relations (« partie-de », « possession stricte », « attribut », etc., voir la section 2.7.3). La relation en jeu est simplement une relation de « spécification » : le complément spécifie l'interprétation du syntagme qu'il complète.

Un problème possible serait un manque d'homogénéité des descriptions complètes proposées par différents annotateurs. Dans le cas des dates, il est facile de définir un langage contrôlé, dans le cas général que nous proposons, cela est beaucoup moins évident.

Cela étant, rien ne dit que cette méthode permettrait d'obtenir une meilleure opérationalité. Ce type de traitement a été utilisé pour le Lancaster Anaphoric Treebank [33], où un type de relation anaphorique est « inferable of-complement », c'est-à-dire qu'un syntagme x est anaphorique par rapport à un syntagme y si y peut être un complément de x introduit par la préposition *of*. Nous n'avons pas connaissance d'une évaluation de l'inter-subjectivité des observations menée dans le cadre du projet Lancaster Anaphoric Treebank.

4.10 Conclusion de la première partie

Nous avons défini dans le premier chapitre la notion de reprise. Cette notion est conçue comme une généralisation de la notion d'anaphore, dans le sens où elle ne met pas en jeu de conditions sur la forme des expressions, mais elle pose aussi des restrictions par rapport à l'anaphore, dans le sens où les relations en jeu sont limitées à des liens sémantiques caractérisés en ayant recours à une relation d'identité. L'intérêt de la généralisation est qu'elle permet de se concentrer sur le résultat du processus d'interprétation des expressions en jeu, plutôt que sur le processus lui-même. L'intérêt de la restriction à un ensemble de liens caractérisés en ayant recours à une relation d'identité est que cette restriction devrait permettre d'aboutir à une meilleure inter-subjectivité des observations.

Dans le chapitre 2, nous avons ensuite présenté une typologie des différents liens de reprise pouvant être observés dans les textes. Avec ces deux premiers chapitres, nous avons spécifié un système d'observation du réel que sont les textes. Il s'agissait de définir ce que nous voulions observer et comment il devait l'être. C'est là la première étape par laquelle doit passer celui qui adopte la démarche des sciences du réel.

Nous reprenons les termes de G. Bès [11, p. 16] :

> [Dans une science du réel,] il faut, au minimum, se donner un système d'observation du réel sur lequel [les hypothèses] portent, système permettant de [les] tester [...]. Ce réel n'est pas préexistant ou immédiatement donné : il faut le convenir, l'expliciter. On ne demande donc pas une science du réel qui traite d'un Réel en majuscule, mais d'un *réel-observé*, c'est-à-dire d'un Réel que l'on a convenu d'observer de telle ou telle manière.

On notera que nous avons développé dans les deux premiers chapitres des définitions et des descriptions que l'on pense indépendantes de toute hypothèse à venir sur ce qui régit le réel que nous voulons observer.

Dans la perspective d'étudier plus avant le réel que nous avons circonscrit, jusqu'à la formulation d'hypothèses qu'il nous faudra alors tester, il faut se poser les questions suivantes. Confrontés aux mêmes données, différents observateurs parviendront-ils aux mêmes observations du réel visé ? Si des écarts entre les observations existent comment les quantifier ? Comment mesurer la validité des hypothèses que nous pourrions être amenés à formuler ?

Ces questions font l'objet des chapitres 3 et 4. Dans le chapitre 3, nous avons proposé de nouveaux critères et mesures d'évaluation pour l'identification des coréférences, répondant par là aux deuxième et troisième questions ci-dessus. On remarquera que les critères proposés, contrairement à certains des critères existants auparavant, permettent de faire abstraction du processus d'interprétation pour ne regarder que le résultat de ce processus. En ce sens, comme la notion de reprise, ils sont indépendants des hypothèses qui peuvent être faites sur le processus d'interprétation lui-même.

Le chapitre 4 aborde la question de l'inter-subjectivité des observations à travers une expérience dans laquelle cinq observateurs devaient noter les différents relations à distance qu'ils observaient entre les expressions de quelques textes. L'enjeu était d'attester que les conditions d'évaluation des hypothèses à venir existent. Les résultats montrent que pour bon nombre de relations, ce n'est pas le cas. Cet échec rend les questions auxquelles l'expérience était censée répondre d'autant plus pertinentes et justifie ainsi la méthode adoptée dans cette première partie de la thèse.

Deuxième partie

Interprétation automatique des expressions pronominales

Chapitre 5

Introduction

L'objectif du travail décrit dans la seconde partie de la thèse est l'implantation d'un système d'interprétation automatique d'un ensemble d'expressions pronominales. L'interprétation de ces expressions constitue un sous-ensemble des phénomènes de reprise. L'implantation d'un système rendant compte de l'ensemble des phénomènes de reprise dépasserait largement le cadre d'une thèse et, si nous avons envisagé un tel système théoriquement dans la première partie, rappelons (voir p. 14) que notre objectif pratique pour la seconde partie était beaucoup plus limité. On notera cependant que le système décrit ici constitue une première étape vers la tâche plus générale d'identification des reprises et que le travail présenté a été conduit dans le respect de la démarche scientifique défendue dans la première partie de la thèse : le système proposé est un système d'hypothèses testables et qui seront effectivement testées.

Le présent chapitre donne les premiers éléments permettant de situer le travail décrit dans la seconde partie de la thèse. On définit dans un premier temps notre objectif (5.1), puis on décrit l'environnement dans lequel le travail a été réalisé, c'est-à-dire essentiellement les outils d'analyse linguistique automatique utilisés (5.2). La section 5.3 décrit la méthodologie adoptée pour parvenir à la définition du système d'interprétation des expressions pronominales retenues.

Enfin, le chapitre se termine par la présentation de l'organisation de l'ensemble de la seconde partie de la thèse.

5.1 Objectif

Notre objectif est de définir et implanter un système d'hypothèses sur l'interprétation de certaines expressions pronominales. La présente section caractérise précisément cet objectif. On définit en premier lieu quelles seront les expressions pronominales que le système devra traiter (5.1.1), puis on définit ce qu'on attend en sortie du système pour ces expressions (5.1.2). L'objectif défini dans la

section 5.1.2 est ensuite caractérisé par la spécification de la « clé » au regard de laquelle le système sera évalué (5.1.3) et d'un prédicat d'évaluation global définissant les conditions que devra remplir la sortie du système pour être jugée parfaitement correcte (5.1.4).

5.1.1 Expressions anaphoriques visées

Expressions pronominales retenues. Le système de règles décrit dans la seconde partie de la thèse vise à spécifier l'interprétation dans les textes des formes pronominales suivantes [1] :

- *il, ils, elle, elles*, que nous appelerons pronoms clitiques sujet ;
- *l', le, la, les*, que nous appelerons pronoms clitiques accusatifs ;
- *lui, leur*, que nous appelerons pronoms clitiques datifs ;
- *lui, elle, eux, elles*, que nous appelerons pronoms disjoints ;
- *son, sa, ses, leur, leurs*, que nous appelerons déterminants posessifs.

Ces formes s'entendent bien évidemment abstraction faite des variations dues aux majuscules (par exemple, on traitera aussi bien les formes *Il* ou *IL*, ou même *-t-il* dans le cas des pronoms interrogatifs, que la forme *il*). Dans la suite de la thèse, nous désignerons ces formes par le terme « expressions pronominales retenues », ou, plus simplement, « expressions pronominales », lorsqu'il n'y aura pas d'ambiguïté sur l'extension donnée au terme.

Ambiguïtés catégorielles. Quatre formes sont ambiguës relativement aux appellations que nous leur donnons : *lui, leur, elle*, et *elles*.

Les formes *elle* et *elles* sont des pronoms disjoints si elles n'occupent pas la fonction de sujet ou si elles occupent cette fonction et sont suivies immédiatement de l'adverbe *aussi*. Par exemple, dans les phrases suivantes

(1) Elle aussi veut venir.

(2) Il a fait cela pour elle.

le pronom *Elle* est un pronom disjoint.

La forme *lui* est un pronom clitique datif si elle occupe la fonction de complément du verbe et est antéposé au verbe (ou postposé si le verbe est à l'impératif) ; elle est un pronom disjoint dans tous les autres cas. Dans la phrase

(3) Il lui parle.

lui est un pronom clitique datif, mais dans la phrase

(4) Lui parle.

où le pronom occupe la fonction de sujet, *Lui* est un pronom disjoint.

[1] Des restrictions sur les conditions dans lesquelles nous spécifierons l'interprétation de ces expressions seront apportées dans la section suivante.

L'ambiguïté de la forme *leur* ne devrait pas poser de problème. Dans la phrase

(5) Ils leur donnent leur congé.

la première occurence de *leur* est un pronom, la seconde un déterminant.

Expressions pronominales exclues. Toute expression que l'on a l'habitude de caractériser comme pronominale et qui n'appartient pas à la liste que nous avons exposée est exclue de notre champ d'observation. Par exemple, étant donné le texte

(6) Reste quatre administrateurs : parmi eux, trois ont démissionné de leur fonction en début d'année.

notre système visera à spécifier l'interprétation de *eux* et *leur* mais pas l'interprétation de *trois*.

Étant donné la phrase

(7) Néanmoins, cette démission n'est pas exclusive d'une radiation, celle-ci étant une sanction disciplinaire.

notre système ne dira rien sur sur l'interprétation de *celle-ci*.

Nous considérons les suites *lui-même, elle-même, eux-mêmes* et *elles-mêmes* comme des unités lexicales atomiques (c'est-à-dire que nous ne les décomposons pas en sous-parties qui seraient par exemple *lui*, - et *même* pour la suite *lui-même*) et ces formes n'appartiennent pas à l'ensemble des expressions anaphoriques dont nous nous proposons de spécifier l'interprétation.

Justification des choix. Le choix de restreindre nos objectifs à l'interprétation du sous-ensemble d'expressions pronominales que nous avons défini est motivé par les considérations suivantes.

En premier lieu, en limitant notre champ d'observation à un ensemble réduit de formes, nous avons voulu faire abstraction de particularités qui pourraient être liées à certaines formes [2]. Notre démarche a consisté à nous concentrer sur un ensemble réduit d'expressions pronominales, pour en rendre compte le mieux possible, avec dans l'idée qu'une fois notre système d'hypothèses défini pour ces expressions, nous pourrions le tester dans un travail futur sur des formes pronominales non prévues au départ et examiner dans quelle mesure nos hypothèses rendent ou non compte de l'interprétation de ces formes.

Par ailleurs, l'exclusion des pronoms relatifs de notre champ d'étude est due au fait que ces formes étaient déjà traitées par l'analyseur syntaxique que nous avons été amenés à utiliser.

[2] Par exemple, les formes telles que *celui-ci, celui-là*, etc., mettent sans doute en jeu une contrainte sur la linéarité du texte qui n'est pas présente pour l'interprétation des pronoms clitiques : *celui-ci* et *celui-là* s'opposent respectivement en ce que l'un indique qu'on fait référence à un être dénoté par une expression plus proche, l'autre qu'on fait référence à un être dénoté par une expression plus lointaine, opposition que l'on ne retrouve pas avec les pronoms clitiques, par exemple.

L'exclusion des pronoms réfléchis (pronom clitique *se* et pronom disjoint *soi*) est motivée par une volonté de nous concentrer sur des phénomènes inter-phrastiques ou du moins potentiellement inter-phrastiques. Les pronoms réfléchis étant presque toujours coréférents avec le sujet du verbe dont ils sont compléments, on peut les considérer comme implicitement résolus par l'analyseur syntaxique. Les calculs à effectuer pour ces expressions seraient très simples et l'inclusion de ces expressions dans les expressions visées conduirait à une amélioration des résultats qui masquerait en partie le réel problème posé par les expressions que nous avons effectivement retenues.

Ces remarques sur les pronoms réfléchis sont également valables pour les formes *lui-même, elle-même,* etc, d'où l'exclusion de ces dernières de l'ensemble des expressions visées par notre système.

Pour terminer cette justification du choix que nous avons fait, il convient de noter que les expressions pronominales que nous avons choisi de traiter, en particulier les pronoms clitiques et les déterminants possessifs, sont de loin les plus fréquentes dans les textes [3], si bien que notre système d'hypothèses concernera une part importante des expressions pronominales apparaissant effectivement dans les textes.

5.1.2 Spécifier l'interprétation des expressions pronominales

DÉFINITION. Notre objectif est de « spécifier l'interprétation » ou de « donner l'interprétation » des expressions pronominales retenues. On entend par là le fait de relier chaque occurence d'une expression pronominale retenue e_i à au moins une expression e_j telle que

(a) e_j est une expression pronominale ou un syntagme nominal qui ne soit pas une coordination de syntagmes nominaux,
(b) e_i et e_j sont coréférentes.

À ces deux conditions s'ajoute la contrainte (c) suivante :

(c) si, pour une expression pronominale e_i, il existe à la fois e_j et e_k qui satisfont les conditions (a) et (b), et telles que e_j appartient à l'ensemble des expressions pronominales dont nous nous proposons de spécifier l'interprétation et e_k n'appartient pas à cet ensemble, alors l'objectif est de relier e_i à e_k et e_j à e_k (et non simplement relier e_i et e_j entre elles).

[3] Si on fait abstraction des pronoms relatifs et réfléchis, qui sont traités par ailleurs par l'analyseur syntaxique. Dans le corpus annoté dans le cadre du projet décrit dans [90], les expressions pronominales retenues ici représentent environ 75 % des reprises, sur un ensemble d'un peu plus de 22 600 reprises (les phénomènes retenus étant les reprises par pronoms, sauf réfléchis et relatifs, déterminants possessifs et syntagme nominaux avec ellipse du noyau). Le chiffre donné ici ne tient pas compte des occurrences du pronom *il* impersonnel.

La phrase suivante illustre la situation décrite dans la condition (c).

(8) En réalité, la banque a plusieurs fois laissé entendre qu'elle était prête à faire son deuil de ce « privilège ».

Dans cette phrase, les expressions *la banque* (e_k) et *elle* (e_j) satisfont les conditions (a) et (b) relativement au déterminant possessif *son* (e_i), mais l'expression *elle* appartient à l'ensemble des expressions pronominales à interpréter. Au final, l'objectif sera de relier *son* à *la banque* et non simplement *son* à *elle*.

C'est la condition (c) qui nous permet de dire que, dans le cas général, nous « spécifierons » l'interprétation des expressions pronominales retenues. En effet, dans la très grande majorité des cas, l'expression à laquelle devra être reliée une expression pronominale sera un syntagme nominal non pronominal, c'est-à-dire une expression plus spécifique que l'expression pronominale (mais pas nécessairement l'expression la plus spécifique possible). Feront exception à cette règle générale les cas où une expression pronominale pourra être reliée à une autre expression pronominale, mais qui n'appartient pas à l'ensemble dont nous voulons spécifier l'interprétation. L'exemple (6) 171, reproduit ici, donne un exemple de cette situation :

(6) Reste quatre administrateurs : parmi eux, trois ont démissionné de leur fonction en début d'année.

Pour ce texte, notre objectif sera de relier *leur* à *trois*, expression qui ne peut pas, à proprement parler, être considérée comme plus spécifique que *leur*.

Nous appelerons l'expression à laquelle doit être reliée une occurence d'une expression pronominale, la « source » de cette expression pronominale.

Restrictions apportées par la définition. La définition de notre objectif exposée ci-dessus pose quelques restrictions qui n'apparaîssent peut-être pas à première vue. Nous les explicitons ici.

La condition (a) pose des restrictions sur le type de source que nous envisageons d' identifier. Sont exclus de notre champ d'observation les cas où un pronom renvoie à une phrase ou proposition, comme dans l'exemple suivant :

(9) Chacun le sait, l'hôtel Matignon use terriblement ses locataires.

où le pronom clitique *le* renvoie à la proposition *l'hôtel Matignon use terriblement ses locataires*. Cela ne concerne que les deux formes *le* et *l'*.

Sont également exclus les cas où une expression pronominale renvoie à plusieurs expressions, cas que nous désignons comme des « reprises à source multiple ». Une expression pronominale e_i interprétée par coréférence est jugée comme une reprise à source multiple si la première source à gauche de e_i est un ensemble de syntagmes nominaux noyau, ou bien, si aucune source n'existe à gauche de e_i, si la première source à droite de e_i est un ensemble de syntagmes nominaux noyau. Une coordination de syntagmes nominaux est considérée comme un ensemble de syntagmes nominaux.

Dans la phrase :

(10) San Paolo et IMI chiffrent les économies obtenues grâce à leur fusion.

le déterminant possessif *leur* est interprété comme dénotant l'ensemble constitué des entreprises San Paolo et IMI. Ces deux sociétés sont désignés par deux syntagmes distincts coordonnés. Notre système d'hypothèses ne visera pas à rendre compte de ce type de phénomène.

Dans le texte suivant :

(11) En dépit des différences importantes entre la position stratégique des courriéristes parlementaires au Québec et celle des chroniqueurs de l'éducation en France, l'objectif de légitimation de leur pouvoir symbolique est commun aux deux groupes. Ils optent d'ailleurs pour des stratégies similaires.

Les expressions *leur*, *[les] deux groupes* et *Ils* dénotent toutes trois l'ensemble constitué des courriéristes parlementaires au Québec et des chroniqueurs de l'éducation en France. L'ensemble constitué des syntagmes *des courriéristes parlementaires au Québec* et *des chroniqueurs de l'éducation en France* est une source possible pour *leur* et *Ils*. Comme cet ensemble de syntagmes est la première source à gauche de *leur*, cette expression pronominale est considérée comme une reprise à source multiple. En revanche, dans le cas de *Ils*, l'expression *aux deux groupes* est une source possible au même titre que l'ensemble de syntagmes nominaux considéré, si bien que cette expression pronominale n'est pas une expression à source multiple.

Il y a dans ce choix un présupposé : si, dans un texte, il est fait référence à un ensemble d'êtres `X` à la fois par un ensemble d'expressions dénotant chacune un sous-ensemble ou un être particulier de `X` (p. ex. *Allianz* et *Ergo*) et par une expression unique (p. ex. *les deux sociétés*), l'expression unique aura plutôt tendance à suivre l'ensemble d'expressions.

Le choix de ne pas rendre compte des reprises à source multiple est lié à une difficulté à représenter ces cas compte tenu, d'une part, de la manière dont la structure syntaxique des phrases est représentée par l'analyseur syntaxique que nous avons utilisé et, d'autre part, du système dans lequel nous avons implanté nos descriptions. de la représentation donnée de la structure syntaxique des phrases et du système dans lequel nous avons implanté nos descriptions (voir la description de l'analyse syntaxique en entrée du système au chapitre 7).

La condition (b) exclut les cas où un pronom est interprété par rapport à sa source avec une relation qui n'est pas une identité de dénotation. Cela concerne les deux formes *le* et *l'*, comme dans la phrase suivante

(12) Jacques est président et Lionel ne l'est pas.

où le pronom *l'* reprend la description *président*. Notons que, que ce soit par la condition (a) ou la condition (b), les reprises avec *le faire* sont également exclues.

Enfin, nous souhaitons bien entendu relier une expression pronominale avec une autre expression dans la mesure où une telle expression existe. Les emplois

impersonnels du pronom *il*, comme dans la phrase suivante, constituent les principaux cas où un pronom ne renvoie à aucune expression.

(13) Il est possible qu'il pleuve.

Autre exemple où un pronom ne renvoie à aucune expression, le clitique *l'* dans la phrase :

(14) Le secrétaire général de la CGT, Louis Viannet, a estimé vendredi que le « diktat » du commissaire européen à la concurrence l'avait « emporté ».

Cette description des différents cas particuliers qui sont exclus de notre champ d'observation nous conduit à une remarque importante : dans une situation idéale, lorsque nous testerons notre système sur un texte quelconque, nous supposerons que les expressions pronominales de ce texte pour lesquelles notre système ne doit pas spécifier d'interprétation ont été identifiées préalablement comme telles.

5.1.3 Données spécifiées par la clé

Pour finir de caractériser notre objectif, nous spécifions ici quelle sera la « clé » au regard de laquelle notre système d'hypothèses devra être testé. Cette spécification est l'occasion de présenter quelques exemples de ce qu'on souhaite obtenir en sortie du système.

Spécification de la clé. Pour un texte T donné, la « clé » est constituée des données spécifiées par l'analyse par un observateur humain des expressions pronominales que notre système tentera d'interpréter, analyse effectuée en conformité avec les définitions et restrictions posées dans les deux sections précédentes.

Pour un texte T, la clé spécifie l'ensemble K_T des chaînes de coréférence du texte, avec la restriction que ces chaînes ne contiennent que des expressions qui satisfont la condition (a) et/ou des expressions pronominales à résoudre. On rappelle ici notre définition des chaînes de coréférence, qui est basée sur une définition préalable des chaînes de référence (voir la section 3.2).

Étant donné un texte T, une chaîne de référence CR_i est caractérisée comme l'ensemble des expressions qui dénotent un être de l'univers de dénotation `o1` donné.

$$CR_i = \{e_i \mid e_i \text{ est une expression dénotant } \texttt{o1} \text{ dans } T\}$$

Une chaîne de coréférence CC_i est une chaîne de référence ayant au moins deux éléments.

Avec K_T la clé pour un texte ou un ensemble de textes T, on a :

$$K_T = \{CC \mid CC \text{ est une chaîne de coréférence dans } T\}$$

où, on le rappelle, on parle de chaîne de coréférence restreintes aux expressions qui satisfont la condition (a) ou sont des expressions pronominales à résoudre.

Nous nous intéresserons plus particulièrement aux chaînes de coréférence de K_T qui contiennent au moins une expression pronominale dont notre objectif est de spécifier l'interprétation. Nous dirons de ces expressions pronominales qu'elles sont les « reprises » de K_T [4]. Soit R_k l'ensemble des reprises de la clé.

La clé spécifie en outre pour chaque chaîne de coréférence une partition en deux sous-ensembles $\texttt{pron}_{CC_i}$ et $\texttt{src}_{CC_i}$, respectivement l'ensemble des reprises visées [5] de CC_i et l'ensemble des sources possibles de CC_i pour les éléments de $\texttt{pron}_{CC_i}$.

$$\texttt{pron}_{CC_i} = \{e_i \in CC_i \mid e_i \in R_k\}$$

$$\texttt{src}_{CC_i} = \{e_i \in CC_i \mid e_i \notin R_k\}$$

Exemples. Soit le texte suivant, qui contient quatre expressions pronominales appartenant à l'ensemble des expressions retenues (ces expressions sont en *italiques* et les diverses occurrences d'une même forme sont distinguées par des indices) :

(15) Robert Panhard, cinquante-deux ans, a été élu hier président de la Chambre des notaires de Paris. Titulaire d'une maîtrise de droit et diplômé de l'Institut des études politiques de Paris, *il*$_1$ débute *sa* carrière comme fondé de pouvoir à la BIMP (Banque Industrielle et Mobilière Privée). Diplômé du notariat en 1979, *il*$_2$ rejoint la société civile professionnelle Dauchez, Kubisa, Panhard, Baffoy et Deneuville. *Il*$_3$ entend désormais valoriser la profession, en constituant un réseau notarial européen, ancré à Paris, renforcer le pôle immobilier et moderniser les études notariales.

Les quatre expressions pronominales à considérer dans ce texte dénotent toutes le même être de l'univers de dénotation, à savoir Robert Panhard. Cet être est par ailleurs dénoté dans ce texte par l'expression *Robert Panhard*.

En nous limitant à la seule chaîne de coréférence qui contienne une expression pronominale, on a la clé suivante :

$$K_{(15)} = \{\ \{\text{Robert Panhard, il}_1\text{, sa, il}_2\text{, Il}_3\}\ \}$$

Appelons CC_1 l'unique élément de $K_{(15)}$. On a la partition suivante :

$$\texttt{src}_{CC_1} = \{\text{Robert Panhard}\}$$

$$\texttt{pron}_{CC_1} = \{\text{il}_1\text{, sa, il}_2\text{, Il}_3\}$$

[4] Le terme « reprise » est donc ici employé avec une acception plus réduite que dans la première partie de la thèse. Il doit être compris comme « reprises qui doivent être interprétées par notre système de résolution ».

[5] Il s'agit des expressions pronominales dont nous nous proposons de spécifier l'interprétation.

Autre exemple. Le texte suivant contient deux expressions pronominales appartenant à l'ensemble des expressions retenues :

(16) Le rapprochement des deux banques est en tout cas défendu par Vincenzo Maranghi [...]. Celui-ci, après avoir vu le Credito Italiano torpiller *son* projet de *le* fusionner avec Banca di Roma et Comit (qui, ensemble, détiennent environ 25 % de Mediobanca), milite aujourd'hui pour l'alliance entre Comit et Banca di Roma.

En nous limitant une nouvelle fois aux seules chaînes de coréférence qui contiennent au moins une expression pronominale, on a la clé suivante :

$$K_{(16)} = \{\ CC_2,\ CC_3\ \}$$

avec

$$CC_2 = \{\text{Vincenzo Maranghi, Celui-ci, son}\}$$

$$CC_3 = \{\text{Credito Italiano, le}\}$$

Pour ces deux chaînes, on a les partitions suivantes :

$$\texttt{src}_{CC_2} = \{\text{Vincenzo Maranghi, Celui-ci}\}$$

$$\texttt{pron}_{CC_2} = \{\text{son}\}$$

$$\texttt{src}_{CC_3} = \{\text{Credito Italiano}\}$$

$$\texttt{pron}_{CC_3} = \{\text{le}\}$$

5.1.4 Prédicat d'évaluation global

À partir des données spécifiées par la clé, on peut définir un prédicat d'évaluation global pour la sortie de notre système, prédicat qui résumera notre objectif.

Pour un texte T, la sortie du système consistera en un ensemble S_T de couples (e_i, A_{e_i}) où e_i est une reprise selon le système et A_{e_i} est l'ensemble des « antécédents possibles » pour e_i, toujours selon le système. L'ensemble des antécédents possibles pour une reprise e_i en sortie du système est un ensemble d'expressions avec lesquelles le système dit e_i coréférente.

La sortie de notre système de résolution des expressions pronominales sera parfaitement correcte si et seulement si :

$$\forall CC_i \in K_T, \forall e \in \texttt{pron}_{CC_i}, \exists A_e, (e, A_e) \in S_T, A_e \neq \emptyset \text{ et } A_e \subset \texttt{src}_{CC_i}$$

et

$$\forall (e_i, A_{e_i}) \in S_T, e_i \in R_k$$

Notons que l'exigence que l'ensemble A_{e_i} des antécédents d'une expression e_i en sortie du système soit inclus dans l'ensemble des expressions *sources* de la chaîne de coréférence à laquelle appartient e_i dans la clé exprime la contrainte (c) de notre définition d'objectif (voir p. 172).

EXEMPLES. Étant donné notre prédicat d'évaluation, la sortie attendue du système pour les deux exemples de la section précédente doit être la suivante.

Pour l'exemple (15), une seule sortie est possible :

$$S_{(15)} = \{ \text{ (il}_1\text{,\{Robert Panhard\})},\text{(sa,\{Robert Panhard\})}, \\ \text{(il}_2\text{,\{Robert Panhard\})},\text{(Il}_3\text{,\{Robert Panhard\}) } \}$$

Pour l'exemple (16), en revanche, trois sorties $S^1_{(16)}$, $S^2_{(16)}$ et $S^3_{(16)}$ sont possibles :

$$S^1_{(16)} = \{ \text{ (le,\{Credito Italiano\})},\text{(son,\{Vincenzo Maranghi, Celui-ci\}) } \}$$

$$S^2_{(16)} = \{ \text{ (le,\{Credito Italiano\})},\text{(son,\{Vincenzo Maranghi\}) } \}$$

$$S^3_{(16)} = \{ \text{ (le,\{Credito Italiano\})},\text{(son,\{Celui-ci\}) } \}$$

Selon notre prédicat d'évaluation, ces trois sorties sont parfaitement correctes. On pourrait être tenté de juger que l'une d'elles est plus intéressante que l'autre (par exemple que $S^2_{(16)}$ est plus intéressante que $S^3_{(16)}$ car *Vincenzo Maranghi* est une expression plus spécifique que *Celui-ci*), mais l'observateur qui spécifie la clé doit résiter à une telle tentation : on ne lui attribue pas la capacité de juger la qualité relative de deux sorties parfaitement correctes.

5.2 Environnement de travail

La présente thèse est le fruit d'un travail effectué dans le cadre d'une Convention industrielle de formation par la recherche (CIFRE) au Centre européen de recherche de Xerox (XRCE). Ce laboratoire de recherche inclut une équipe spécialisée dans le traitement automatique des langues dont l'objectif est de développer des applications effectives, c'est-à-dire qui aillent au-delà du stade de la simulation ou de l'illustration théorique. En particulier, a été développé à XCRE un analyseur syntaxique, appelé XIP (pour « Xerox Incremental Parser », [3]), analyseur que nous avons été amenés à utiliser. L'objet de cette section est de présenter dans leurs grandes lignes cet outil et la chaîne de traitements dans laquelle il est utilisé. Notre propre système de résolution des pronoms s'inscrit dans cette chaîne de traitements.

Les différentes étapes de l'analyse d'un texte sont représentées sur la figure 5.1 page suivante par les rectangles pleins à droite. Au centre, en gras, le nom des outils informatiques utilisés pour accomplir ces différentes tâches (NTM et XIP). Les rectangles en pointillés à gauche représentent les sources d'informations utilisées par les outils informatiques. Les quatre premières étapes permettent d'obtenir l'analyse syntaxique du texte en entrée, qui constitue à son tour l'entrée de notre système de résolution des pronoms. Nous les évoquons ici brièvement et en simplifiant certains aspects des outils ou processus décrits.

5.2.1 Analyse morphologique

La première étape consiste en une segmentation du texte en unités lexicales et analyse morphologique de ces unités. Ces deux tâches sont effectuée en une seule étape grâce à l'outil NTM (pour « Normalization, Tokenization, Morphology »). Cet outil, conçu par Salah Aït-Mokhtar, utilise comme données des automates à états finis (AEF), qui sont eux-mêmes définis par des expressions régulières.

L'analyse morphologique d'une unité lexicale non ambiguë consiste en l'association à cette unité d'un lemme et d'ensemble d'étiquettes qui encodent des informations sur le nombre, le genre, la personne ou encore la catégorie syntaxique de l'unité lexicale analysée [6]. Lorsqu'une unité lexicale est ambiguë, le système retourne plusieurs analyses. Par exemple, étant donné en entrée la phrase de la figure 5.2, le processus d'analyse morphologique produira la sortie présentée figure 5.3. Dans cette sortie, une ligne correspond à une analyse possible pour une unité lexicale donnée ; un interligne sépare la ou les analyses d'une unité lexicale de celle(s) d'une autre unité. Une ligne comprend trois parties : à gauche l'unité lexicale analysée telle qu'elle figure dans le texte en entrée, au centre le lemme associé à l'unité lexicale dans l'analyse considérée, et à droite une suite d'« étiquettes » encodant les informations évoquées plus haut.

Sans trop entrer dans les détails, l'exemple de la figure 5.3 nous permet d'illustrer le type d'information fournie en sortie l'analyse morphologique. Pour chaque analyse, l'étiquette la plus à droite indique à quelle partie du discours (pronom, verbe, adjectif, nom, etc.) appartient l'unité lexicale analysée. Par exemple, la forme *ferme* peut être un verbe, dont le lemme est *fermer* ou un adjectif, ou un nom, ou encore un adverbe (p. ex. *On s'ennuie ferme.*). Les étiquettes `+P1`, `+P2` et `+P3` donnent l'information sur la personne, les étiquettes `+SG` et `+PL` l'information sur le nombre, les étiquettes `+Masc`, `+Fem` et `+InvGen` (« invariable en genre ») l'information sur le genre. Pour les pronoms, les étiquettes `+Nom` et `+Acc` donnent l'information sur le cas (*il* est un pronom nominatif, *le* est un pronom accusatif). Pour les verbes, l'analyse morphologique donne l'information sur le temps et le mode (`+SubjP` = subjonctif présent, `+IndP` = indicatif présent, `+Imp` = impératif) et le type d'auxiliaire avec lequel la forme du passé composée est construite.

[6] Cette liste n'est pas exhaustive.

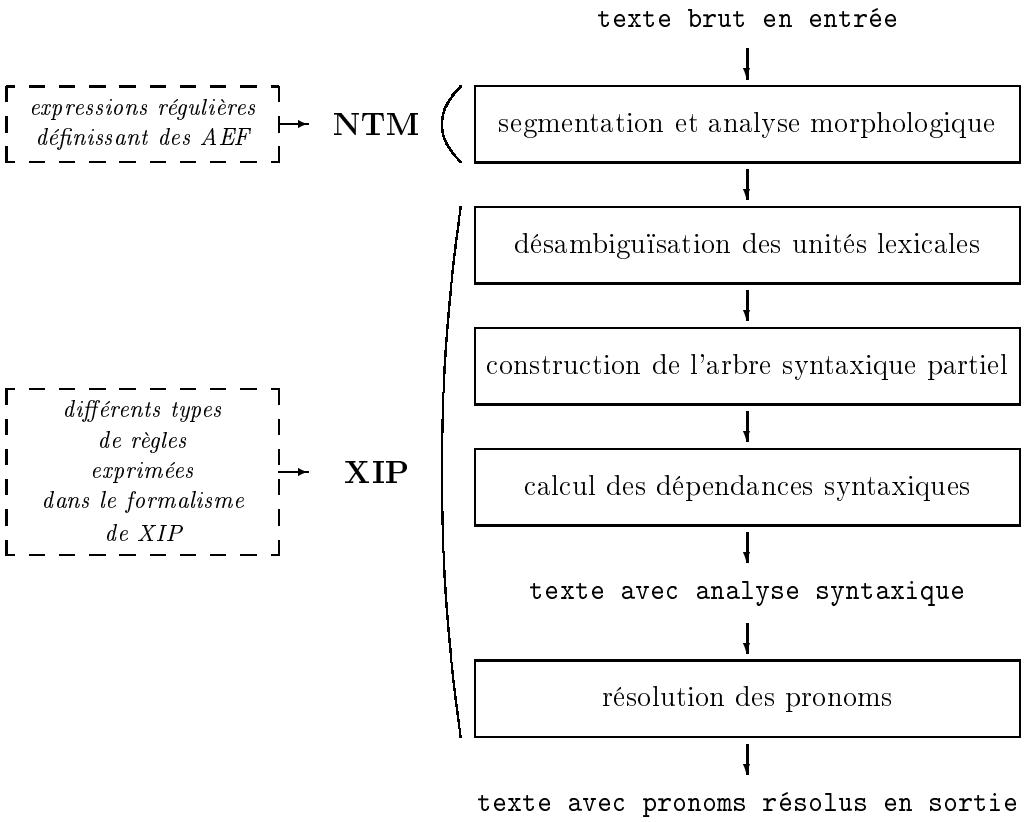

FIG. 5.1 – Architecture du système

À ces informations morphologiques *stricto sensu*, s'ajoutent d'une part une information indiquant le type de compléments sous-catégorisés par les noms, verbes et adjectifs (p. ex. les étiquettes `+se`, `+aSN` et `+SN` pour le verbe *fermer*), d'autre part une étiquette à valeur sémantique pour les noms pouvant être utilisés pour dénoter un lieu ou une personne au sens large (c'est-à-dire une personne physique, un groupe de personnes, une persone morale). Le nom *ferme* peut ainsi être utilisé pour désigner un lieu (étiquette `+Lieu`).

5.2.2 Analyse syntaxique

Les deuxième, troisième et quatrième étapes, réunies, produisent l'analyse syntaxique du texte. Ces trois étapes sont effectuées grâce à l'outil XIP, qui dans tous les cas prend comme source d'informations des règles formulées dans un formalisme propre à l'outil lui-même (par exemple des règles analogues aux règles de réécriture classique). L'outil XIP lui-même est un analyseur générique, c'est-à-dire un système qui lit un ensemble de règles, ou formules, décrivant une langue donnée et, étant donné un texte en entrée, produit l'analyse de ce texte telle que spécifiée par les règles ou formules. Différentes règles peuvent bien entendu être définies pour différentes langues. À cet égard, les trois étapes décrites ici sont celles par lesquelles passe l'analyse de textes en français, étant donné le système de règles définies par les linguistes de XRCE pour cette langue. La présence de ces trois étapes n'est pas une exigence due à l'outil XIP lui-même.

Dans un premier temps, la sortie de l'analyseur morphologique est désambiguïsée, c'est-à-dire que pour chaque unité lexicale une seule analyse est retenue. Étant donné l'analyse morphologique de la figure 5.3, la désambiguïsation a pour résultat la sortie présentée figure 5.4. Parmi les huit lectures possibles de la forme *ferme*, est retenue celle qui en fait un verbe à la troisième personne du présent de l'indicatif. La forme *les*, quant à elle, est analysée comme un déterminant défini pluriel qui peut être masculin ou féminin.

Dans un deuxième temps, le système construit un arbre syntaxique *partiel*, puis, dans un troisième temps, calcule des « dépendances » entre les nœuds de cet arbre. Nous ne nous attardons pas ici sur ces deux étapes ; les données qu'elles produisent seront largement décrites au chapitre 7.

5.2.3 Système de résolution des pronoms

Le système XIP qui est utilisé pour effectuer l'analyse syntaxique telle que nous l'avons brièvement décrite s'est révélé suffisamment générique pour qu'on l'utilise avec profit pour notre tâche de résolution des pronoms. Ce que nous appelons ici « système de résolution des pronoms » n'est donc pas un outil informatique spécifique dédié à la résolution des pronoms, mais un système de formules exprimées dans le formalisme propre au système XIP.

```
Il ferme les volets.
```

FIG. 5.2 – Phrase en entrée du système

```
Il       il       +Nom+Masc+SG+P3+Pron

ferme    fermer   +se+aSN+SN+avoir+SubjP+SG+P1+Verb
ferme    fermer   +se+aSN+SN+avoir+SubjP+SG+P3+Verb
ferme    fermer   +se+aSN+SN+avoir+Imp+SG+P2+Verb
ferme    fermer   +se+aSN+SN+avoir+IndP+SG+P1+Verb
ferme    fermer   +se+aSN+SN+avoir+IndP+SG+P3+Verb
ferme    ferme    +dansSN+avecSN+surSN+InvGen+SG+Adj
ferme    ferme    +Lieu+Fem+SG+Noun
ferme    ferme    +Adv

les      le       +Acc+InvGen+PL+P3+Pron
les      le       +InvGen+PL+Def+Det

volets   volet    +Masc+PL+Noun

.        .        +SENT
```

FIG. 5.3 – Analyse morphologique pour la phrase de la figure 5.2

```
Il       il       +Nom+Masc+SG+P3+PC

ferme    fermer   +se+aSN+SN+avoir+IndP+SG+P3+Verb

les      le       +InvGen+PL+Def+Det

volets   volet    +Masc+PL+Noun

.        .        +SENT
```

FIG. 5.4 – Analyse morphologique désambiguïsée

L'outil XIP est donc utilisé à la fois pour l'analyse syntaxique et pour la résolution des pronoms, ce qu'indique sur la figure 5.1 la parenthèse à droite du nom **XIP**, parenthèse qui englobe ces différentes étapes. Notre travail a donc consisté à fournir une nouvelle source d'information au système XIP.

Dans la chaîne de traitement, cette source d'information est destinée à être utilisée après l'analyse syntaxique. L'ensemble des formules que nous avons définies, conjointement aux sources d'informations utilisées dans l'analyse morphologique et l'analyse syntaxique, constitue notre hypothèse sur l'interprétation des pronoms.

5.3 Méthodologie

L'objectif du travail présenté dans cette seconde partie de la thèse est de formuler un système d'hypothèses — ou plutôt *une* hypothèse, puisque le système doit être vu comme un tout — sur l'interprétation des expressions pronominales spécifiées dans la section 5.1.1. Nous présentons ici les aspects principaux de la méthodologie qui a conduit à la définition de ce système.

5.3.1 Un système implanté en machine

Dans le contexte d'un travail effectué en entreprise, notre objectif était de définir un système d'hypothèses effectivement implantable et implanté en machine, système qui puisse ainsi être utilisable pour une application de traitement automatique des langues telle que l'indexation ou la recherche d'informations. Autrement dit, le système implanté doit pouvoir fonctionner sur des textes réels, avec un temps de calcul raisonnable.

Cette exigence a pour conséquence qu'il devient préférable pour nous de formuler notre hypothèse sur la base d'une information qui peut effectivement nous être fournie automatiquement. C'est pourquoi nous avons choisi de travailler sur la sortie de l'analyseur syntaxique du français développé à XRCE, plutôt que sur un corpus annoté, quitte à ce que des imperfections viennent pertuber les résultats et leur évaluation. On fait l'hypothèse que ces perturbations seront assez faibles.

Par ailleurs, l'implantation effective et complète (c'est-à-dire analyse syntaxique incluse) du système en machine a aussi pour intérêt de permettre une évaluation de l'hypothèse qu'il traduit sur une échelle significative, la quantité d'information utilisée se révélant vite trop importante pour une évaluation non automatique.

5.3.2 Validité statistique de notre hypothèse

À notre connaissance, personne, à l'heure actuelle, n'a été capable de décrire complètement ce qui régit l'interprétation des pronoms, dans quelque langue que

ce soit. Dans ce contexte, nous ne rechercherons donc pas pour notre hypothèse une validité absolue. Plus modestement, nous souhaitons qu'elle soit valide pour une quantité importante de pronoms, en d'autres termes qu'elle ait une « validité statistique ». La notion de « quantité importante » est bien évidemment floue : il s'agira de juger si les résultats sont ou non satisfaisants.

5.3.3 Étude de corpus

L'exigence d'un système implanté en machine et la limitation de notre objectif à une validité statistique nous ont conduit à adopter comme support de découverte de notre système d'hypothèses l'étude de corpus. On se donne donc au départ un corpus d'étude, dont l'observation nous permet de définir notre système, et ce système sera ensuite évalué sur un nouveau corpus, dit « corpus d'évaluation ».

Le corpus d'étude et le corpus d'évaluation sont caractérisés par leur source et leur thème. En l'occurence, ces corpus sont pour nous des articles du journal d'informations économiques *La Tribune*, dont le thème est plus spécifiquement la banque et l'assurance. Les articles utilisés pour le test d'opérationnalité présenté au chapitre 4 donnent un exemple du type de textes sur lesquels nous avons travaillé (voir l'annexe A).

En définissant notre système d'hypothèses à partir de l'observation d'un corpus et en l'évaluant sur un corpus de même type, il y a un risque que le système d'hypothèses soit très spécifique, d'une part, à ce qui serait en fait un idiolecte, en l'occurrence celui des auteurs des articles de *La Tribune*, d'autre part, au sujet traité dans ces articles, que l'on peut caractériser globalement comme des informations sur les entreprises du monde de la finance. Nous sommes conscients de ce risque. Notre approche a consisté à définir un système qui soit éventuellement spécialisé avec dans l'idée que ce degré de spécialisation pourra être ensuite évalué par application du système sur de nouveaux corpus de sources et de domaines différents.

5.4 Plan de la deuxième partie

La seconde partie de la thèse est organisée comme suit.

Dans le chapitre 6, nous décrivons les différentes sources d'informations qui peuvent être utilisées pour interpréter les pronoms et nous présentons les principaux systèmes d'interprétation automatique développés à ce jour.

Les chapitres 7 et 8 donnent les éléments qui permettront de comprendre le système d'interprétation des expressions pronominales que nous avons défini et qui sera décrit aux chapitres 9, 10 et 11. Le chapitre 7 décrit l'entrée du système, qui consiste en une analyse syntaxique du texte sous la forme d'un arbre syntaxique partiel et de dépendances entre les nœuds de cet arbre. Le chapitre 8

décrit le formalisme du système XIP, système dans lequel sont exprimées nos hypothèses sur l'interprétation des pronoms.

Le chapitre 9 décrit l'organisation globale du système. On dit quelles sont les expressions pronominales dont nous nous proposons de spécifier l'interprétation et quel sera le critère d'évaluation global. Est ensuite présentée la structure du système, que se décompose en cinq types de formules correspondant à cinq étapes du processus d'analyse.

Les chapitres 10 et 11 décrivent dans le détail les différentes formules qui traduisent nos hypothèses sur l'interprétation des expressions pronominales retenues. Ces formules sont regroupées en deux grands ensembles : les « règles », d'une part, objet du chapitre 10, et les « préférences » d'autre part, objet du chapitre 11.

Le chapitre 12 présente l'évaluation détaillée du système. Globalement, le système donne une interprétation correcte pour 75 % des expressions pronominales visées, évaluation effectuée sur un ensemble de 417 expressions apparaissant dans un recueil d'articles de presse traitant du domaine de la finance. Nous verrons dans le chapitre 12 quelle est l'apport des différents composants du système à ce résultat.

Chapitre 6

Approches du problème

Avant d'en venir à la description de notre système d'interprétation des expressions pronominales que nous avons retenues, nous rendons compte dans le présent chapitre, d'une part, des différentes sources d'information qui peuvent être utilisées pour l'interprétation des pronoms, d'autre part, des principaux systèmes d'interprétation des pronoms qui ont été implantés et évalués.

L'information pertinente pour l'interprétation des expressions pronominales peut être répartie en quatre types : une information de nature syntaxique (section 6.1), des informations de nature sémantique, soit lexicale, avec les restrictions de sélection (section 6.2), soit de nature pragmatique (section 6.3), et une information relevant de la cohérence et de la structure du discours (section 6.4). Ces différents types d'informations sont présentés successivement dans les quatre sections suivantes.

La section 6.5 décrit les principaux algorithmes de résolution des pronoms implantés à ce jour et la section finale met brièvement notre propre système en perspective par rapport à ces algorithmes.

6.1 Syntaxe

Les travaux effectués dans le domaine de la théorie syntaxique, en particulier ceux de Chomsky [21] et Reinhart [77], ont mis à jour les contraintes structurelles qui, au niveau de la phrase, régissent l'emploi et l'interprétation des pronoms. Ces contraintes s'articulent autour des notions de « liage » et de « c-commande ». La présente section expose les grandes lignes de ces contraintes [1].

[1] Pour cette section, notre principale source d'information a été la synthèse de A. Zribi-Hertz, *L'anaphore et les pronoms* [96]. Chomsky et Reinhart ne sont bien évidemment pas les seuls à avoir travaillé sur les questions évoquées ici, mais ce sont sur leurs définitions que nous nous appuyons. Voir notre ouvrage de référence [96] pour une étude plus complète.

Par ailleurs, nous tenons pour acquis qu'en règle générale, un pronom s'accorde en genre et en nombre avec son antécédent, si bien que nous n'évoquerons pas cette source d'information

6.1.1 La théorie du liage

La théorie du liage [21] définit trois classes d'expressions anaphoriques selon les contraintes qui pèsent ou non sur leur interprétation [2] :

- les expressions pronominales qui doivent être liées dans leur domaine local (type A, il s'agit des pronoms dits « réfléchis » ou « réciproques », p. ex. le pronom *se*) ;
- les expressions pronominales qui sont libres (c'est-à-dire ne peuvent pas être liées) dans leur domaine local (type B, il s'agit des pronoms non réfléchis, p. ex. le pronom *le*) ;
- les expressions anaphoriques non pronominales, qui ne peuvent jamais être liées (type C, p. ex. un syntagme nominal démonstratif).

Cette catégorisation correspond à trois « principes » de la théorie.

La notion de liage est définie en relation avec celle de c-commande :

> Si deux expressions e_i et e_j sont coréférentes et si e_i « c-commande » e_j, alors e_j est liée par e_i.

La notion de c-commande est définie comme une relation entre deux nœuds dans une représentation syntaxique de la phrase sous forme d'arbre. Soient deux nœuds A et B et X le premier nœud à ramifications qui domine A [3].

> Le nœud A c-commande le nœud B si
> 1. A ne domine pas B, et inversement, et si :
> 2.a. soit X domine B,
> 2.b. soit X est immédiatement dominé par un nœud X' de même type catégoriel que X et ce nœud X' domine B.

Le domaine local d'une expression e est défini comme le plus petit syntagme qui contient à la fois e et un sujet. Étant donné une représentation en arbre de la structure de la phrase, un syntagme X est la suite d'unités lexicales (autrement dit de nœuds terminaux de l'arbre) dominées par le nœud X. La notion de sujet dans cette définition est assez étendue : d'une part, elle inclut le sujet non réalisé effectivement d'un infinitif (on a alors une unité lexicale vide parmi les nœuds terminaux de l'arbre), d'autre part, un syntagme prépositionnel Y complément d'un nom N ou un possessif Y déterminant un nom N sont considérés comme des sujets si le nom N exprime une propriété de l'être dénoté par Y (si N et Y

dans le présent chapitre.

[2] On évoque ici la théorie du liage dans la version simplifiée pour les données du français, telle que présentée dans [96, p. 117]. En particulier, cette version ne fait pas usage de la notion de gouvernement.

[3] La définition donnée ici est reprise de [96, p. 59], avec cependant une formulation légèrement différente, qui rectifie une erreur de formulation sur le point 2.b dans [96] (le nœud X' ne doit pas être de même type catégoriel que le nœud A mais de même type catégoriel que le nœud X).

constituent une « prédication », voir [96, p. 107 et 115]).

Les exemples suivants, repris de [96], illustrent l'application de la théorie du liage aux expressions anaphoriques de type A. Les expressions pronominales et leurs antécédents sont en italiques et marqués par la lettre i en indice. Les crochets délimitent le domaine local de l'expression pronominale considérée. Dans ces exemples, l'expression pronominale *ne peut pas* être interprétée autrement que comme coréférente avec l'antécédent spécifié. Dans l'exemple (1b), le symbole ϵ représente l'unité lexicale vide sujet de l'infinitif.

(1) a. [*Pierre*$_i$ *se*$_i$ photographiera].

b. Pierre$_i$ a promis à Marie [de ϵ_i *se*$_i$ photographier].

c. Pierre connaît [*leur*$_i$ mépris *les uns pour les autres*$_i$].

Inversement, dans les deux exemples suivants, structurellement équivalents à (1a) et (1b), respectivement, les expressions pronominales de type B ne peuvent être coréférentes avec les expressions marquées par $*i$ en indice ; cela parce que les expressions pronominales seraient alors *liées* (les antécédents potentiels considérés c-commandent les pronoms en question).

(2) a. [*Pierre*$_{*i}$ *le*$_{*i}$ photographiera].

b. Pierre$_{*i}$ a promis à Marie [de ϵ_{*i} *le*$_{*i}$ photographier]

Une expression de type B peut être liée par son antécédent si celui-ci est en dehors de son domaine local. Dans la phrase suivante, le syntagme *Pierre* c-commande le pronom *le*, mais est en dehors du domaine local de *le*.

(3) *Pierre*$_i$ pense que [quelqu'un *le*$_i$ photographiera].

Une expression de type B peut être coréférente avec une expression de son domaine local si celle-ci ne la c-commande pas. C'est le cas de *Pierre*, par rapport à *le*, dans la phrase suivante :

(4) [La sœur de *Pierre*$_i$ *le*$_i$ photographiera].

Notons que dans les exemples (3) et (4), la coréférence entre les expressions pronominales et les antécédents envisagés est possible, mais non nécessaire (c'est-à-dire que le pronom *le* dans ces deux phrases pourrait dénoter une personne autre que Pierre). Pour les expressions pronominales de type B, la théorie du liage spécifie seulement des contraintes de non-coréférence.

6.1.2 Contraintes de non-coréférence

Dans la mesure où les expressions pronominales visées par le système d'interprétation automatique que nous nous proposons de définir (voir la section 5.1.1, page 170) sont toutes des expressions de type B, les contraintes de non-coréférence exposées dans la section précédente constituent l'élément de la théorie du liage qui sera susceptible de nous intéresser.

La contrainte de non-coréférence posée par la théorie du liage pour les expressions pronominales qui nous intéressent, contrainte que nous nommerons « contrainte de non-liage », peut être reformulée comme suit (nous employons dorénavant le terme « expression pronominale » pour désigner une expression visée par notre système, donc une expression de type B) :

> CONTRAINTE DE NON-LIAGE. Une expression pronominale ne peut pas être coréférente avec une expression appartenant à son domaine local et qui la c-commande.

À cette contrainte, s'ajoutent deux autres contraintes non explicitées dans la définition des classes d'expressions anaphoriques énoncée plus haut : d'une part, la contrainte de c-commande [96, p. 56] :

> CONTRAINTE DE C-COMMANDE. Une expression anaphorique ne peut pas c-commander son antécédent.

d'autre part, la contrainte sur l'antécédent quantifié [96, p. 89] :

> CONTRAINTE SUR L'ANTÉCÉDENT QUANTIFIÉ. Un antécédent quantifié doit c-commander l'expression qui l'anaphorise.

La contrainte de c-commande est illustrée par l'exemple suivant :

(5) Chez *Pierre*$_{*i}$, *il*$_{*i}$ fume le narguilé.

On a pour cette phrase la structure suivante :

$$[_S\ [_{SP}\ \text{Chez Pierre}\ _{SP}],\ [_{S'}\ [_{SN}\ \text{il}\ _{SN}]\ [_{SV}\ \text{fume le narguilé}\ _{SV}]\ _{S'}]\ _S].$$

Le nœud S domine immédiatement le nœud S', ce dernier est le premier nœud à ramifications dominant le pronom *il*, et le nœud S domine *Pierre*. Le pronom *il* c-commande donc l'expression *Pierre* et ne peut être coréférent avec elle.

Les syntagmes nominaux quantifiés sont des syntagmes tels que *chacun*, *personne*, *quelqu'un*, *aucun homme*, *chaque homme*, etc. T. Reinhart a remarqué que les conditions de reprise de ce type de syntagme sont plus contraintes et a donc formulé la contrainte sur l'antécédent quantifié. Elle est motivée par le contraste suivant :

(6) a. L'infirmière qui soigne *Pierre*$_i$ avec amour *l'*$_i$adore.

b. L'infirmière qui soigne *chacun*$_{*i}$ avec amour *l'*$_{*i}$adore.

En (6a), le pronom peut être coréférent avec *Pierre* ; il ne peut pas être coréférent avec *chacun* en (6b).

Les contraintes que nous avons présentées ici sont dépendantes de la manière dont est représentée la structure syntaxique de la phrase [4]. Nous verrons par

[4] De fait, dans les travaux théoriques sur la syntaxe, la définition des contraintes et celle de

la suite (au chapitre 7) que l'analyse syntaxique qui nous est donnée en entrée n'a pas du tout la structure requise pour qu'on puisse exprimer directement ces contraintes. Nous serons donc amenés à les exprimer d'une manière différente.

Par exemple, la contrainte de c-commande exclut en particulier de nombreux cas où un pronom pourrait avoir pour antécédent une expression qui le suit, plutôt qu'une expression qui le précède : ainsi un pronom sujet du verbe principal d'une proposition p c-commande tous les syntagmes nominaux qui sont dominés par le nœud qui représente p, or ceux-ci sont en général des expressions qui *suivent* le pronom sujet. La situation est similaire pour un pronom clitique objet, qui est immédiatement dominé par le nœud représentant le groupe verbal et c-commande donc tous les syntagmes nominaux du groupe verbal. N'ayant pas moyen, faute de l'arbre syntaxique *ad hoc*, d'implanter la contrainte de c-commande telle qu'elle est formulée plus haut, nous formulerons une règle qui explicitera positivement les cas où un pronom peut renvoyer à une expression qui le suit (voir la règle décrite p. 316).

Autre exemple avec la contrainte de non-liage. Nous n'exprimerons pas cette contrainte en termes de c-commande, mais nous formulerons une règle disant qu'un pronom clitique non réfléchi qui est complément d'un verbe ne peut être coréférent avec le sujet de ce verbe (voir la règle décrite p. 333 ; le cas traité est illustré par l'impossibilité d'une coréférence entre *le* et *il* dans *il le voit*).

On notera que de manière générale, nous ne serons pas en mesure de mettre à jour de façon systématique et complète la correspondance entre nos règles et les différentes contraintes décrites dans la présente section (c'est-à-dire de montrer que nos règles excluent bien tout ce que les contraintes en question ici excluent). Pour cela, il nous faudrait une correspondance entre une représentation syntaxique sur laquelle les contraintes peuvent être exprimées et la représentation syntaxique qui nous est donnée et nous ne disposons pas d'une telle correspondance.

On peut cependant signaler qu'il n'y a rien dans notre système qui ressemble à la contrainte sur l'antécédent quantifié. Les couples <expression pronominale, antécédent quantifié> susceptibles d'être exclus par cette seule contrainte étant très rares dans les textes.

6.2 Restrictions de sélection

Un facteur qui intervient probablement dans l'interprétation des expression pronominales est ce qu'on appelle les « restrictions de sélection ». Dans une première sous-section, nous présentons cette notion à partir d'un article de M.

la structure syntaxique sont intimement liées (voir, par exemple, la représentation de l'accord sujet-verbe par un nœud AGR dans la théorie du liage, telle que décrite dans [96, p.137]).

Saiz-Noeda et M. Palomar [80], qui proposent précisément de l'utiliser pour l'interprétation des pronoms [5]. La seconde sous-section est consacrée à la description de quelques expériences et résultats obtenus dans des systèmes effectivement implantés.

6.2.1 Idée générale

L'exemple suivant permettra une première approche du problème. Il semble que rien, structurellement, dans la phrase donnée en (7), n'empêche que le déterminant possessif *sa* soit interprété comme faisant référence à Patrick Careil ; or il ne fait aucun doute que ce déterminant fait référence à la SMC.

(7) Les audits réalisés à la demande de Patrick Careil, qui préside la SMC le temps de sa privatisation, préconisent « un nettoyage de type paille de fer ».

L'idée fondamentale des restrictions de sélection appliquées à l'interprétation des pronoms est que c'est parce que les objets qu'on privatise sont typiquement des sociétés, et non des êtres humains, que l'interprétation de *sa* dans l'exemple (7) se fait sans ambiguïté aucune. On dit que le nom *privatisation* « sélectionne » de préférence un complément de type « société ».

Saiz-Noeda et Palomar [80] décrivent un projet d'utilisation des restrictions de sélection pour l'interprétation des pronoms sujets et objets en espagnol. Nous nous limitons ici à la description de leur méthode pour les pronoms sujets, les choses étant égales, *mutatis mutandis* pour les pronoms objets.

On suppose que pour chaque pronom sujet un ensemble d'antécédents possibles est identifié sur la base de critères touchant à la linéarité du texte et à la morpho-syntaxe (p. ex. l'ensemble des syntagmes nominaux de la phrase où apparaît le pronom et ceux de la phrase précédente, qui s'accordent avec le pronom, sont des antécédents possibles ; c'est le type de méthode utilisé dans les approches automatiques que nous verrons plus loin). L'objectif est d'identifier l'antécédent correct dans cet ensemble.

On se donne deux ontologies O_n pour les noms et O_v pour les verbes. Les ontologies de Saiz-Noeda et Palomar sont des arbres dont les feuilles sont des noms (resp. des verbes) et les nœuds des « caractéristiques sémantiques » des noms (resp. des verbes). Un nœud n_i qui domine un nœud n_j ou une feuille f_i est une caractéristique sémantique plus générale que n_j ou celle qui est exprimée par f_i (p. ex. « substance » domine (entre autres) « nourriture », qui domine « fruit » qui domine « banane »). Les auteurs citent WordNet comme une ontologie analogue à la leur.

[5] Les auteurs envisage un traitement des pronoms en espagnol, mais les restrictions de sélection sont *a priori* utilisables pour n'importe qu'elle langue. Signalons que Y. Wilks [94] a été un des précurseurs en matière d'utilisation des restrictions de sélection, dès 1973.

On a par ailleurs un ensemble de « relations de compatibilité sujet-verbe » (R_{sv}). Une relation R_{sv} est une paire $[n, v]$ où n est un nom ou une caractéristique sémantique de O_n et v est un verbe ou une caractéristique sémantique de O_v compatible avec n. Les auteurs donnent l'exemple de la paire [*fruit*,*être mûr*] qui se traduit par « la caractéristique sémantique *fruit*, en tant que sujet, est compatible avec le verbe *être mûr* ».

Muni de ces données, lorsqu'on a dans un texte

- un verbe V_i dont le sujet est un pronom P,
- un ensemble A d'antécédents possibles pour P,

et dans l'ensemble des relations R_{sv}

- une relation $[n, V_i]$ ou $[n, SV_i]$, avec SV_i une caractéristique sémantique qui domine V_i et n une caractéristique sémantique qui domine l'un des noms parmi les antécédents possibles de P,

alors on sélectionne ledit nom comme antécédent de P.

Exemple. Dans le texte suivant, on a trois antécédents possibles pour le pronom *Il* : *le singe, l'arbre*, et *un fruit*.

(8) Le singe est monté sur l'arbre pour prendre un fruit. Il n'était pas mûr.

Parmi les trois antécédents possibles, *un fruit* est sélectionné car il est ou est dominé par une caractéristique sémantique compatible avec le verbe *être mûr* alors que ce n'est pas le cas des deux autres antécédents.

6.2.2 Utilisation effective des restrictions de sélection

Saiz-Noeda et Palomar indiquent que, à la date de l'article dont nous avons résumé le propos dans la section précédente, l'implantation et l'évaluation de leur méthode sont en projet. Nous n'avons pas connaissance d'un nouvel article qui présenterait des résultats effectifs. Quelques auteurs ont néanmoins utilisé et/ou évalué l'apport des restrictions de sélection, ou d'une information comparable, pour la tâche d'interprétation des pronoms. Dans la mesure où l'information sur les restrictions de sélection est particulièrement coûteuse à produire (il s'agit d'examiner tour à tour des milliers de mots), celles-ci sont en général utilisées en mettant en jeu des techniques d'acquisition automatique.

Une expérience de Dagan et Itai

I. Dagan et A. Itai [26] rendent compte d'une expérience visant à valider l'utilisation de statistiques sur les cooccurrences de termes dans différentes structures syntaxiques pour la résolution des pronoms. Les cooccurrences de termes en question, dites « patrons de collocation » sont équivalentes à des restrictions

de sélection, à cette différences près que ces dernières sont en générale exprimées pour des *classes* de termes.

Le principe de l'expérience est illustré par l'exemple suivant. La phrase en (9) contient deux occurrences du pronom *it* ; la première est sujet du verbe *collect* et la seconde objet de ce même verbe.

(9) They know full well that the companies held tax money aside for collection later on the basis that the government said it_1 was going to collect it_2.

Les auteurs envisagent trois antécédents possibles pour chacune de ces deux occurrences : *money, collection* et *government*.

L'observation du corpus révèle que *government, money* et *collection* apparaissent comme sujet du verbe *collect* respectivement 198, 5 et 0 fois. Les noms *government* et *collection* n'apparaissent jamais comme objet du verbe *collect* dans le corpus, contrairement au nom *money* (149 fois). À partir de ces données, on peut déduire que le pronom *it* sujet de *collect* renvoie très probablement à *government* et que le pronom *it* objet de *collect* renvoie à *money*.

Étant donné un corpus de taille importante, les auteurs enregistrent des « patrons de cooccurrence » pour la relation sujet-verbe et la relation verbe-objet, c'est-à-dire le nombre d'occurrence des paires (n, v), où n est un nom sujet du verbe v, et des paires (v, n), où n est objet du verbe v. Rappelons que les données extraites sont plus spécifiques que les restrictions de sélection puisqu'elles concernent des termes, alors que les restrictions de sélection sont en général exprimée relativement à des classes de termes (p. ex. l'ensemble des termes décrivant un être animé).

Sont ensuite extraites du même corpus 59 phrases contenant une occurrence du pronom *it* pouvant renvoyer à plusieurs antécédents après application des contraintes morpho-syntaxiques (tout ceci étant effectué manuellement). On a donc 59 paires $(p, A(p))$, où p est une occurrence du pronom *it* et $A(p)$ est l'ensemble des antécédents possibles pour le pronom p. Pour un pronom p quelconque, $A(p)$ contient toujours au moins deux éléments, le but étant d'évaluer l'apport des patrons de cooccurrence collectés.

Si, pour une paire (p, a_i), où a_i est un antécédent possible de p, il existe un nombre significatif (plus de 5) de patrons de cooccurrence reliant une occurrence du nom a_i et une occurrence du verbe dont dépend p, alors cet antécédent est retenu comme antécédent possible pour p.

Dans 21 cas sur 59 (36 %), les statistiques collectées ne sont pas suffisamment significatives pour être utilisées (moins de 5 patrons ont été collectés pour chacun des antécédents). Pour 33 des 38 cas restant, l'antécédent correct figure parmi les antécédent retenus (soit une précision de 87 % selon ce critère), l'ensemble des antécédents retenus se réduisant à un élément dans 18 cas.

Ces résultats montrent à la fois l'intérêt des cooccurrences de termes, et plus largement des restrictions de sélection, et leurs limites : si le taux de précision

est intéressant, les cooccurrences de termes extraites par Dagan et Itai, malgré leur spécificité, conduisent à des erreurs d'interprétation en quantité significative (dans 8,5 % des cas (5/59) l'antécédent correct n'est pas retenu).

Autres mesures de l'apport des restrictions de sélection

T. Nasukawa [65], dont nous décrivons le système de résolution des pronoms plus loin (section 6.5.8), utilise également les patrons de cooccurrence apparaissant dans le texte précédant un pronom p pour interpréter ce pronom. Sur un échantillon de 84 pronoms, un patron a été rencontré dans 26,2 % des cas. Dans tous les cas, l'utilisation des patrons produit une interprétation correcte.

L'apport de la technique développée par Dagan et Itai a été évaluée en tant que complément de l'algorithme de Lappin & Leass ([55], décrit plus loin dans la section 6.5.3), algorithme utilisant au départ essentiellement des informations de nature syntaxique et de mise en focus des référents par références multiples. Sans les patrons de cooccurrence, le système obtient un taux de succès [6] de 86 % ; avec les patrons de cooccurrence, ce score s'élève à 89 %. L'utilisation des patrons de cooccurrence seulement, conduit à interpréter 51 % des pronoms seulement, avec une précision de 79 %. Lappin & Leass concluent de cette expérience que les informations utilisées par leur algorithme sont plus pertinentes que les patrons de cooccurrences extraits du corpus.

Les résultats obtenus par Dagan et Itai, Lappin & Leass et Nasukawa montrent que, si l'observation d'exemples choisis tels que les exemples (7) ou (8) de la section 6.2.1 met en lumière l'intérêt des restrictions de sélection de façon assez convaincante, celles-ci ne peuvent constituer la seule source d'information pour l'interprétation des pronoms, au-delà des contraintes morpho-syntaxiques. Ils mettent également en lumière l'intérêt d'une évaluation d'un système effectivement implanté par rapport à une évaluation « manuelle ». Dans le domaine de la sémantique lexicale, la quantité d'information à gérer est telle qu'il est impossible, lorsqu'on observe un cas particulier, d'envisager ce cas particulier dans le contexte plus large de la base lexicale. Il est ainsi facile d'oublier une ambiguïté lexicale (p. ex. étant donné *il marche*, avec *il* interprété comme dénotant une personne, on dira que *marcher* sélectionne un sujet dénotant un être animé, en oubliant que ce même verbe peut être employé dans le sens de « fonctionner ») ou un usage métaphorique (p. ex. on pourrait remplacer la seconde phrase de (8) par *Il est mûr pour quitter sa mère*), ou, du moins, il est impossible de quantifier l'influence de tels facteurs sur le processus de résolution.

[6] On parle de taux de succès, plutôt que de rappel et précision, lorsqu'un système donne une réponse pour tous les pronoms et n'est pas susceptible d'identifier à tort une expression comme pronominale. Dans ce cas, le nombre de réponse en sortie est égal au nombre de pronoms à résoudre dans la clé, et le rappel et la précision sont donc égaux.

6.3 Pragmatique

Les contraintes de non-coréférence et les restrictions de sélection, si elles permettent de décrire un certains nombre de cas, ne suffisent pas à rendre compte de l'interprétation des expressions pronominales. Le processus d'interprétation peut mettre en jeu des inférences d'un ordre que nous qualifions de « pragmatique ». T. Winograd [95] en a donné un exemple souvent cité, que nous adaptons ici :

(10) a. Les élus ont refusé aux grévistes le droit de manifester car ils craignaient la violence.

b. Les élus ont refusé aux grévistes le droit de manifester car ils prônaient la violence.

Les deux phrases sont structurellement semblables et les verbes *craindre* et *prôner* sélectionnent le même type de sujet. En (10a), le pronom *ils* est interprété comme dénotant les élus ; en (10b), il est interprété comme dénotant les grévistes. Globalement, le raisonnement qui conduit à l'une ou l'autre de ces interprétation repose sur les points suivants :

- dans les deux cas, la seconde proposition est interprétée comme une justification du refus exprimé dans la première ;
- il est supposé qu'une manifestation serait susceptible d'engendrer la violence ;
- qui craint la violence refuse ce qui peut l'engendrer (10a) ;
- qui prône la violence ne refuse pas ce qui peut l'engendrer (10b).

Dans le même ordre d'idée, Carbonnel et Brown [16] évoquent ce qu'ils appellent les contraintes de pré-condition/post-condition. Dans le texte suivant, le pronom *Il* ne peut être coréférent avec *Pierre* car l'action exprimée par la première phrase implique que c'est Jacques qui a la pomme (et donc peut la manger), et non Pierre.

(11) Pierre a donné un pomme à Jacques. Il l'a mangé.

Si les exemples de Winograd et de Carbonnel et Brown peuvent paraître artificiels, on n'en rencontre pas moins des cas analogues dans les corpus. Dans le texte suivant, extrait de notre corpus d'étude, le déterminant possessif *son* est interprété comme dénotant René Barberye.

(12) Les manoeuvres d'approche pour succéder à *René Barberye*$_i$ auraient d'ailleurs déjà commencé. Raymond Douyère notamment, auteur du rapport sur la réforme du groupe, qui sert de base à la réflexion de Bercy, se verrait d'ailleurs bien prendre *son*$_i$ siège, dit-on un peu partout.

En théorie, une autre interprétation pourrait être que ce déterminant dénote Raymond Douyère, mais il serait étrange de dire d'une personne qu'elle se verrait bien prendre un siège qu'elle a déjà.

L'information nécessaire à l'interprétation des expressions pronominales présentées dans ces exemples paraît à l'heure actuelle hors de portée des systèmes de

traitement automatique des langues. Si Hobbs [45] a proposé une formalisation des connaissances requises et un algorithme utilisant ces informations pour la résolution des pronoms, aucune implantation n'a été effectuée et seuls quelques exemples illustrent ce que pourrait être la solution.

Notre propre système d'interprétation des expressions pronominales, comme, à notre connaissance, les autres systèmes effectivement implantés à l'heure actuelle, ne fera pas usage d'information d'ordre pragmatique.

6.4 Cohérence et structure du discours

Il arrive que les contraintes syntaxiques, sémantiques et pragmatiques ne permettent pas de déterminer de manière non ambigüe l'antécédent d'une expression pronominale dont l'interprétation paraît pourtant assez claire à un observateur humain. Un quatrième type d'information joue un rôle dans l'interprétation des pronoms, que l'on peut caractériser globalement comme un ensemble de facteurs relevant de la cohérence et de la structure du discours.

La façon dont la structure locale ou globale d'un discours affecte l'interprétation des expressions pronominales a été beaucoup étudiée. Nous présentons ici en détail ce qui nous paraît être la théorie dominante en ce qui concerne la structure locale du discours : la théorie du centrage. Cette théorie intègre et développe un certain nombre de travaux plus anciens (voir [40]) ; ceux-ci ne seront pas présentés ici dans la mesure où les idées principales sont présentes dans la théorie du centrage [7].

Dans la section 6.4.2, nous présentons la « théorie de veines », qui met à jour une corrélation entre l'interprétation des expressions référentielles et la structure globale du discours.

6.4.1 La théorie du centrage

La théorie du centrage a été développée par B. Grosz, A. Joshi et S. Weinstein à partir de 1983. Cette théorie vise à décrire la relation entre la cohérence locale d'un discours et l'usage des expressions référentielles. Les auteurs avancent que « certaines entités mentionnées dans un énoncé sont plus centrales que d'autres et que cette propriété impose des contraintes sur l'usage de différents types d'expressions référentielles par un locuteur. » [39, p. 203]

La théorie du centrage a donné lieu à des discussions et développements en quantité considérable, dont nous ne pouvons rendre compte ici. Nous nous limi-

[7] On notera simplement que les quatre types d'information caractérisés dans les quatre premières sections du présent chapitre se retrouvent dans l'algorithme proposé par C. Sidner pour l'interprétation des pronoms [83]. Après avoir défini une première règle « naïve », Sidner note que celle-ci « peut être révisée pour inclure des critères relevant de la syntaxe, de la sémantique et de la pragmatique, ainsi que des critères reposant sur les caractéristiques du discours ». Dans [83], Sidner s'intéresse précisément au quatrième aspect.

terons à une présentation de la théorie, puis à la description de deux algorithmes de résolution des pronoms basés sur cette théorie : celui de Brennan et al. [13] et celui de Strube et Hahn [84].

Exposé de la théorie

Les données principales sur lesquelles est exprimée la théorie du centrage sont les « énoncés » d'un segment de discours et les « centres » de ces énoncés [8].

Un discours est composé d'une suite d'énoncés (*utterrances*). Ce qui doit être considéré comme un énoncé est sujet à discussion ; pour simplifier, nous adoptons ici la position de [39], où un énoncé est simplement une phrase, simple ou complexe [9]. On note U_i un énoncé quelconque, U_{i-1} l'énoncé qui précède U_i.

Les *centres* (*centers*) d'un énoncé sont « les entités qui servent à lier un énoncé aux autres énoncés du segment de discours qui le contient » [39, p. 208]. Les centres ne sont pas des expressions mais des « objets sémantiques » qu'on peut assimiler aux entités évoquées par l'énoncé [10]. Pour un énoncé U_i, on distingue l'ensemble des *forward-looking centers*, noté $C_f(U_i)$, et un *backward-looking center*, noté $C_b(U_i)$.

Les *forward-looking centers* $C_f(U_i)$ d'un énoncé U_i sont l'ensemble des entités évoquées par l'énoncé. Cet ensemble est partiellement ordonné, d'une manière censée représenter leur importance relative dans l'énoncé. La hiérarchie définie par cet ordre est l'élément crucial dans la théorie. L'ensemble des facteurs déterminant cet ordre sont jugés dans [39, p. 210] comme restant à déterminer complètement ; une hiérarchie faisant usage des fonctions syntaxiques est cependant proposée, qui privilégie le syntagme sujet sur les syntagmes objets, ces derniers étant privilégiés par rapport aux syntagmes occupant une autre fonction syntaxique :

$$\text{sujet} < \text{objet(s)} < \text{autres fonctions}$$

Le *backward-looking center* d'un énoncé U_i est défini en relation avec l'énoncé U_{i-1}. Étant donné l'ensemble des entités de $C_f(U_{i-1})$ qui sont « réalisées » [11] dans U_i (autrement dit l'intersection de $C_f(U_{i-1})$ et $C_f(U_i)$), le *backward-looking center* de U_i ($C_b(U_i)$) est l'élément de cet ensemble qui occupe la position la plus haute dans la hiérarchie de $C_f(U_{i-1})$.

À partir des données présentées jusqu'à présent, sont définis quatre types

[8] Cet exposé de la théorie du centrage est basée sur l'article récapitulatif de 1995 [39] et sur le chapitre introductif de [93].

[9] Kameyama [48] propose une définition plus fine de la notion d'énoncé, dans laquelle une phrase complexe (contenant plusieurs propositions) est décomposée en plusieurs énoncés.

[10] Voir cette autre définition des centres : « the set *forward-looking centers*, $C_f(U_i, D)$ represents the discourse entities evoked by an utterance U_i in a discourse segment D. » [93, p. 3]

[11] Pour une exposition de la notion de réalisation, voir [39]. On peut considérer le terme « entité réalisée » comme globalement équivalent à « entité évoquée ».

	Dans la hiérarchie de $C_f(U_i)$,	
	$C_b(U_i)$ est le plus haut	$C_b(U_i)$ n'est pas le plus haut
$C_b(U_i) = C_b(U_{i-1})$	CONTINUE	RETAIN
$C_b(U_i) \neq C_b(U_{i-1})$	SMOOTH-SHIFT	ROUGH-SHIFT

TAB. 6.1 – Types de transitions entre énoncés dans la théorie du centrage.

de transitions [93, p. 6] [12] entre deux énoncés U_{i-1} et U_i (voir le tableau 6.1). Ces quatre transitions sont définies en fonction du fait que les *backward-looking centers* respectifs de U_{i-1} et U_i sont identiques ou non [13] et de la hiérarchie des $C_f(U_i)$.

Deux règles complètent la théorie [39, p. 214] : pour chaque énoncé U_i dans un segment de discours D consistant en un ensemble d'énoncés $U_1, \ldots, U_n$,

1. Si un élément de $C_f(U_{i-1})$ est réalisé par un pronom dans U_i, alors c'est aussi le cas de $C_b(U_i)$.
2. Les *séquences* de type CONTINUE sont préférées aux *séquences* de type RETAIN, qui sont préférées aux *séquences* de type SHIFT [14].

À ce stade de la présentation, un exemple semble indispensable. On reproduit ici celui de [39, p. 6] ; on a deux discours, (13) et (14), commençant par les deux mêmes énoncés mais qui diffèrent par le troisième. En (13c), le pronom *He* est interprété comme dénotant Jeff et en (14c) comme dénotant Dick. Après chaque énoncé sont données les informations sur les centres (C_b et C_f) et transitions (T).

(13) a. Jeff helped Dick wash the car.
C_b : *indéfini* / C_f : [JEFF, DICK, CAR] / T : *indéfinie*

b. He washed the windows as Dick waxed the car.
C_b : JEFF / C_f : [JEFF, WINDOWS, DICK, CAR] / T : CONTINUE

c. He soaped a pane.
C_b : JEFF / C_f : [JEFF, PANE] / T : CONTINUE

(14) a. Jeff helped Dick wash the car.
C_b : *indéfini* / C_f : [JEFF, DICK, CAR] / T : *indéfinie*

b. He washed the windows as Dick waxed the car.
C_b : JEFF / C_f : [JEFF, WINDOWS, DICK, CAR] / T : CONTINUE

c. He buffed the hood.
C_b : DICK / C_f : [DICK, HOOD] / T : SMOOTH-SHIFT

[12] Trois, à l'origine, dans [39].

[13] Les relations CONTINUE et RETAIN apparaissent aussi si U_{i-1} ne contient pas de C_b (ce qui se produit quand l'énoncé est le premier du segment de discours).

[14] Ceci pour les trois types de transition de la théorie originale, où les transitions de type SMOOTH-SHIFT et ROUGH-SHIFT ne sont pas distinguées.

D'après la théorie du centrage, le discours de l'exemple (14) serait moins cohérent, c'est-à-dire plus « coûteux » à comprendre, que celui de l'exemple (13) : un locuteur interprètera au premier abord le pronom *He* en (14c) comme dénotant Jeff et sera en quelque sorte contraint de revenir sur sa décision en interprétant complètement la phrase.

Selon une communication personnelle de A. Joshi à M. Strube et U. Hahn, rapportée dans [84, p. 315], « la théorie de centrage n'était pas à l'origine destinée à servir de base à la résolution des expressions anaphoriques. » De fait, *stricto sensu*, la théorie du centrage ne vise pas à spécifier l'interprétation des expressions anaphoriques, mais à porter un jugement sur la plus ou moins grande difficulté qu'il y a à interpréter tel ou tel discours ou à interpréter un même discours de telle ou telle manière.

Cependant, si on suppose que les textes, ou plus généralement les discours, produits effectivement sont cohérents, alors la théorie du centrage peut être utilisée pour spécifier l'interprétation des expressions anaphoriques.

Un algorithme fondé sur la théorie du centrage

Brennan et al. [13] ont été les premiers, en 1987, à présenter un algorithme de résolution des pronoms fondé sur la théorie du centrage. La hiérarchie des C_f pour un énoncé est la suivante :

$$\text{sujet} < \text{objet} < \text{objet2} < \text{autres fonctions sous-catégorisées} < \text{adjoints}$$

Les auteurs notent que cette hiérarchie correspond souvent à l'ordre des expressions dans la phrase anglaise.

Par ailleurs, Brennan et al. introduisent deux modifications à la théorie originale. Ces modifications sont, d'une part, la distinction entre SMOOTH-SHIFT et ROUGH-SHIFT telle que présentée dans le tableau 6.1, d'autre part une reformulation de la règle 2 ci-dessus en :

2′. Une transition de type CONTINUE est préférée à une transition de type RETAIN, qui est préférée à une transition de type SMOOTH-SHIFT, qui est préférée à une transition de type ROUGH-SHIFT.

Là où la théorie originale exprimait des préférences pour des *séquences* de transitions, la règle de Brennan et al. se limite à des préférences sur les transitions seules.

Cela étant, les auteurs proposent un algorithme de résolution des pronoms : les transitions possibles sont d'abord calculées en respectant les contraintes d'accord en nombre et genre des expressions coréférentes, puis filtrées en fonction des contraintes de non-coréférence et des règles de la théorie du centrage, puis classées selon la hiérarchie de la règle 2′. Est alors retenue l'interprétation qui met en jeu la transition la plus haute dans la hiérarchie (voir [13, p. 158] pour le détail de l'algorithme).

Aucune évaluation de l'algorithme n'est donnée par Brennan et al. Signalons que cet algorithme est *a priori* restreint à l'interprétation des pronoms qui renvoient à une expression de la phrase précédente, puisque les énoncés sont définis comme étant les phrases dans la théorie du centrage originale.

Un algorithme fondé sur le « centrage fonctionnel »

Un autre algorithme de résolution des pronoms fondé sur la théorie du centrage a été proposé par M. Strube et U. Hahn [84]. Les auteurs révisent cependant considérablement la théorie originale, en particulier les principes d'ordonnancement des *forward-looking centers*. Plus que les fonctions grammaticales, c'est la « structure d'information fonctionnelle » qui est importante, la structure d'information fonctionnelle consistant, dans sa version la plus simple, à distinguer les expressions qui dans un énoncé dénotent un être supposé connu de l'interlocuteur (*hearer-old*) ou non (*hearer-new*). Une entité est dite *hearer-old* si elle est dénotée par une reprise ou par un nom propre sans apposition ou relative qui caractérise l'identité de l'entité dénotée par le nom propre. Dans les autres cas, une entité est dite *hearer-new* [15].

La hiérarchie des *forward-looking centers* d'un énoncé U_i est définie à partir de la distinction *hearer-old*/*hearer-new*, d'une part, et en fonction de l'ordre des expressions de U_i, d'autre part. Soient x et y deux expressions de U_i dénotant deux entités distinctes X et Y, respectivement :

- Si X est *hearer-old*, et Y est *hearer-new*, alors X est préférée à Y.
- Si X et Y sont toutes deux *hearer-old* ou toutes deux *hearer-new*, et si x précède y, alors X est préférée à Y.

Étant donné cette hiérarchie, Strube et Hahn définissent un « algorithme de base » pour la résolution des pronoms [84, p. 316] :

1. Si un pronom est rencontré dans un énoncé U_i, tester les élément de $C_f(U_{i-1})$ dans l'ordre spécifié par la hiérarchie des C_f jusqu'à ce qu'un élément satisfasse les critères morpho-syntaxiques, les critères de liage et les critères de compatibilité sémantique [16] requis. Cet élément est choisi comme antécédent du pronom.

[15] En ce qui concerne les noms propres, les auteurs donnent les deux exemples suivants.

(15) A defiant Winnie Madikizela Mandela testified [...] today [...].

(16) "He was an underevaluated person all his life," said Marianne Kador, a social worker for Selfhelp Community Services [...].

En (15), l'entité dénotée par *A defiant Winnie Madikizela Mandela* est *hearer-old* (c'est-à-dire qu'elle est censée être connue de l'interlocuteur) ; en (16), l'entité dénotée par *Marianne Kador* est *hearer-new* (la description fournie par l'apposition indique que la personne est *a priori* inconnue de l'interlocuteur).

[16] Les critères de compatibilité sémantique étant ceux qui sont posés par les restrictions de sélection.

2. Lorsque l'énoncé U_i est complètement lu, calculer $C_b(U_i)$ et générer l'ensemble ordonné des $C_f(U_i)$

Notons que même s'il est fait référence au calcul d'un *backward-looking center* dans l'algorithme, celui-ci n'intervient absolument pas dans le calcul permettant d'aboutir à l'interprétation des pronoms [17]. Cette absence, ainsi que l'absence de référence à une typologie des transitions entre énoncés, constituent deux autres différences notables avec la théorie du centrage originale.

Les auteurs ont effectué une évaluation manuelle de cet algorithme sur des textes anglais et allemand contenant respectivement 576 et 619 pronoms personnels ou déterminants possessifs de troisième personne. Ils ont également évalué l'algorithme de Brennan et al. sur les mêmes données. Dans cette évaluation, l'anaphore interne à la phrase est prise en compte aussi bien que l'anaphore externe à la phrase, cela à suivant le découpage des phrases complexes en énoncés proposé par Kameyama [48]. Les critères de compatibilité qui doivent être satisfaits entre le pronom et son antécédent sont strictement ceux de l'étape 1 de l'algorithme ; aucune autre connaissance du monde n'est supposée.

L'algorithme de Brennan et al. obtient un taux de succès de l'ordre de 75 %, celui de Strube et Hahn un taux de succès de l'ordre de 82 %. Ces chiffres semblent indiquer que le fait qu'un antécédent potentiel dénote une entité supposée connue de l'interlocuteur est un critère plus pertinent pour l'interprétation d'un pronom que la fonction syntaxique de cet antécédent (indirectement, l'ordonnancement des C_f d'un énoncé par la position des expressions qui les dénotent est probablement très proche de celui qui serait obtenu par la hiérarchie de Brennan et al., mais ce critère est secondaire chez Strube et Hahn).

D'après Strube et Hahn, une des raisons pour lesquelles leur système donne de meilleurs résultats est le fait que le caractère *hearer-old* d'une entité est défini relativement à une situation plus large que ne le sont les C_b dans la théorie du centrage : ces derniers sont caractérisés par le fait qu'il y a une reprise renvoyant à l'énoncé précédent, alors que les entités sont *hearer-old* simplement si elles sont censées être connues de l'interlocuteur, c'est-à-dire supposées existantes dans une situation dont la connaissance est partagée par le locuteur et l'interlocuteur.

Un autre aspect selon nous important du travail de Strube et Hahn est que ces deux auteurs ont montré qu'une simplification notable de la théorie du centrage, où seule est maintenue l'idée d'une hiérarchisation des entités évoquées dans un énoncé, est susceptible de produire de meilleurs résultats pour la tâche d'interprétation des pronoms.

Pour terminer, signalons que d'un point de vue technique, l'algorithme de Strube et Hahn pose quelques problèmes. M. Poesio et R. Vieira [70] ont mon-

[17] La notion de C_b est seulement utilisée par Strube et Hahn dans une évaluation de leur algorithme en terme de coût de traitement (fréquence de tel ou tel type de transitions dans les textes).

tré qu'il n'est pas acquis que la catégorisation des syntagmes nominaux définis comme dénotant une entité *hearer-old* ou *hearer-new* soit complètement intersubjective. Par ailleurs, à supposer que la distinction *hearer-old/hearer-new* soit opérationnelle, l'obtention de cette information par des moyens automatiques constitue probablement un problème au moins aussi complexe que la résolution des pronoms elle-même.

6.4.2 La théorie des veines

D. Cristea, N. Ide et L. Romary [25] proposent ce qu'ils appellent la « théorie des veines », qu'ils considèrent une généralisation de la théorie du centrage, dans le sens où la théorie des veines est susceptible de s'appliquer sur l'ensemble d'un discours et non seulement localement. Comme la théorie du centrage, la théorie des veines « n'est pas un modèle de la résolution d'anaphores » [25, p. 285] ; nous l'évoquons ici parce qu'un aspect de notre système d'interprétation des pronoms pourra être mis en relation avec certaines idées de la théorie des veines.

Données initiales

La théorie des veines prend pour point de départ une description de l'organisation du discours selon la théorie des structures rhétoriques (RST) de Mann et Thompson [56]. Un texte, dans la théorie des veines, est représenté comme un arbre binaire dont les nœuds terminaux sont des segments de texte (des « unités de discours ») et les nœuds non terminaux représentent des relations entre ces segments. Une polarité est établie entre les fils d'une relation, polarité qui caractérise au moins un nœud comme étant le *noyau* et les autres comme étant les *satellites* du noyau en question. L'unité de discours correspondant au noyau exprime un propos essentiel dans l'intention du locuteur et les satellites expriment une information qui vise à accroître la compréhension mais n'est pas essentielle.

Parmi les exemples donnés figure le texte du haut de la figure 6.1 page 204 (repris de [46]). Chaque ligne constitue une unité de discours, identifiée par une lettre en début de ligne. L'analyse en termes de structures rhétoriques est donnée dans le bas de la figure.

L'étiquette d'un nœud contient en capitales les unités de discours dominées par le nœud et en petites capitales entre crochets la veine associée au nœud. L'étiquetage d'une branche par la lettre « s » indique que le nœud inférieur est un satellite de son nœud frère. Par exemple, les unités de discours B et D sont satellites des unités A et C, respectivement. L'unité complexe EFG est satellite de l'unité complexe ABCD. Il peut ne pas y avoir de relation satellite-noyau entre deux nœuds frères. Les deux nœuds sont alors considérés comme noyaux. Dans l'arbre de la figure 6.1, c'est le cas pour les nœuds F et G, et pour les nœuds ABCDEFG et HIJ.

A. Michael D. Casey, a top Johnson & Johnson manager, moved to Genetic Therapy Inc., a small biotechnology concern here,
B. to become its president and chief opperating officer.
C. Mr. Casey, 46 years old, was president of J&J's McNeil Pharmaceutical subsidiary,
D. which was merged with another J&J unit, Ortho Pharmaceutical Corp., this year in a cost-cutting move.
E. Mr. Casey succeeds M. James Barrett, 50, as president of genetic Therapy.
F. Mr. Barrett remains chief executive officer
G. and becomes chairman.
H. Mr. Casey said
I. he made the move to the smaller company
J. because he saw health care moving toward technologies like the company's gene therapy products.

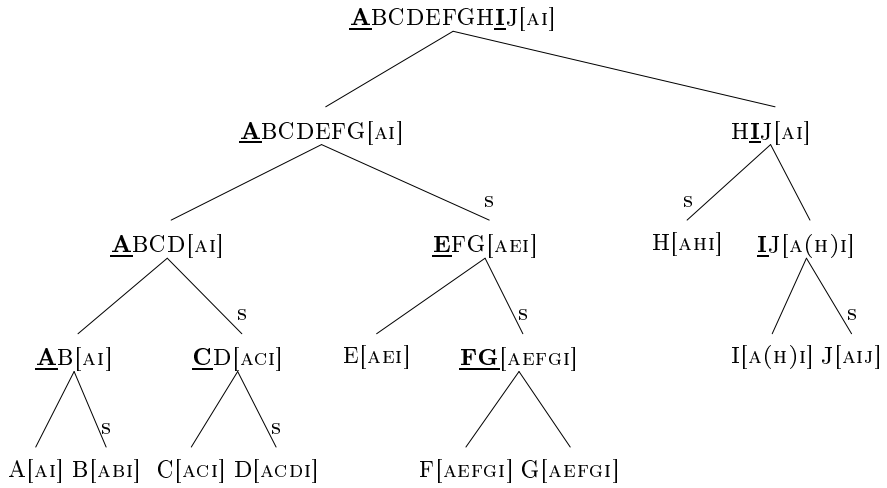

FIG. 6.1 – Un texte et sa structure d'après la théorie des veines.

Étant donné un arbre tel que celui présenté sur la figure 6.1, la théorie des veines définit la notion de « tête » comme suit : la tête d'un nœud terminal est le nœud lui-même et la tête d'un nœud non terminal est la tête de son nœud fils qui est noyau, ou l'union des têtes de ses nœuds fils qui sont noyaux. Les têtes des nœuds non terminaux sont marquées en gras dans la figure 6.1.

À chaque nœud correspond une « veine ». Les veines peuvent être vues comme des ensembles de nœuds terminaux. Soient les trois fonctions suivantes :

- *mark(x)*, qui marque chaque élément de l'ensemble x ;
- *seq(x,y)*, qui concatène les éléments de x avec ceux de y ;
- *simpl(x)*, qui élimine de x les symboles marqués.

La veine du nœud racine est sa tête. La veine d'un nœud n (notée $V(n)$) dont le nœud parent a pour veine v, est calculée comme suit :

1. si n est un nœud noyau
   a. et si n a un nœud frère gauche qui est son satellite et a pour tête h, alors $V(n) = seq(mark(h),v)$;
   b. sinon, $V(n) = v$
2. si n est un nœud satellite ayant pour tête h
   a. et si n est le fils gauche de son nœud parent, alors $V(n) = seq(h, v)$
   b. sinon, $V(n) = seq(h,simpl(v))$

Les veines des différents nœuds sont indiquées sur la figure 6.1 par des petites capitales entre crochets. Les éléments d'une veine qui sont marqués sont placés entre parenthèses.

La dernière notion définie par la théorie des veines est celle de « domaine d'accessiblité ». Le domaine d'accesssibilité d'un nœud terminal est l'ensemble constitué des éléments de sa veine qui le précèdent dans le texte et du nœud lui-même. Par exemple, le domaine d'accessibilité du segment de discours H est constitué des segments A, H et I.

Conjectures

À partir des données initiales de la théorie, Cristea et al. formulent deux conjectures.

La première dit que les « références » à partir d'une unité de discours donnée ne sont possibles que dans son domaine d'accessibilité.

Notons que l'emploi des termes « référence » ou « expression référentielle » dans [25] et [46] n'est pour nous pas très clair. Ils semblent parfois équivalents à « anaphore » ou « expression anaphorique » [18], mais un exemple est par ailleurs

[18] Voir, par exemple, cet extrait de [46] : « The fundamental intuition underlying VT is that the distinction between nuclei and satellites constrains the range of referents to which anaphors can be resolved ; in other words, the nuceleus-satellite distinction induces a domain of referential

donné (dans [46]) où il est question de l'« antécédent » de l'expression référentielle *Mr. Clinton*, qui se trouve être une autre occurrence de l'expression *Mr. Clinton*, un cas qu'il n'est pas d'usage, à notre connaissance, de qualifier d'anaphore. Il semble que si on veut que la premier conjecture de Cristea et al. soit vraiment significative, il faille supposer qu'une « référence » est une relation orientée entre une expression e_i et une autre expression e_j telle que e_j est plus spécifique ou au moins aussi spécifique que e_i, mais ce point n'est pas explicite dans les articles que nous avons consultés.

À la base de la seconde conjecture, se trouve l'idée que la théorie du centrage peut s'appliquer non plus à des énoncés successifs, mais en fonction du domaine d'accessibilité des différentes unités de discours. La conjecture est que cette nouvelle approche donnera lieu à un plus grand nombre de transitions préférées (selon la hiérarchie des transitions de [13]) par rapport à ce qui serait obtenu avec la théorie du centrage initiale.

Les auteurs de la théorie des veines effectuent une première étape vers la validation de ces deux conjectures par une analyse de corpus.

Cristea et al. [25, p. 285] précisent que « la théorie des veines n'est pas un modèle de la résolution de l'anaphore, mais les domaines d'accessibilité qu'elle définit sont un moyen de contraindre la résolution de l'anaphore. L'hypothèse fondamentale de la théorie est qu'une référence inter-unités est possible seulement si les deux unités sont structurellement reliées l'une à l'autre, même si elles sont distantes l'une de l'autre dans le texte. De plus, les références inter-unités renvoient de préférence à une unité noyau, plutôt qu'à une unité satellite, reflétant en cela l'intuition que les noyaux expriment l'idée principale du locuteur. »

L'utilisation de la théorie des veines pour contraindre la résolution des anaphores ne nous semble pas aller de soi, cela parce qu'il est probable que l'interprétation des expressions anaphoriques elle-même intervient dans la détermination de la structure rhétorique d'un discours.

L'unité de discours I dans le texte de la figure 6.1 contient trois expressions référentielles, qui sont toutes les trois anaphoriques (*he, the move to the smaller company* et *the smaller company*). Il est strictement impossible de déterminer comment cette unité de discours s'intègre dans la structure globale sans interpréter ces trois expressions et il est probable que ce soit précisément le fait que ces trois expressions dénotent toutes les trois des êtres évoqués dans l'unité de discours A qui détermine le rattachment des unités HIJ à l'unité A.

Plus qu'une structure indépendante à partir de laquelle on peut exprimer des hypothèses sur l'interprétation des expressions référentielles, Cristea et al. ont, à notre avis, mis à jour une simple corrélation entre les structures rhétoriques et les

accessibility for each referential expression » (l'expression *in other words* invite à considérer *anaphor* et *referential expression* comme deux expressions équivalentes). Il est aussi question de « résolution des expressions référentielles », d'« antécédent d'une expression référentielle », de « co-références *pointant* en dehors de leur unité de discours ».

liens de coréférences. L'hypothèse fondamentale de Cristea et al., telle qu'elle est citée plus haut, dit que si une référence inter-unités est possible, alors les deux unités sont structurellement reliés. Pour que la théorie des veines ait un pouvoir prédictif sur l'interprétation des expressions référentielles, il faudrait s'assurer que les relations entre unités de discours sont toujours établies sur la base de conditions autres que l'existence d'une référence entre les deux segments.

Malgré les réserves que nous avons émises, un aspect de notre système d'hypothèses sur l'interprétation des expressions pronominales n'est pas sans rappeler la notion de noyau et de satellite définie dans le théorie des structures rhétoriques. Les noyaux et les satellites se distinguent par le fait que les noyaux sont supposés contenir une information essentielle et les satellites une information secondaire, non essentielle. Au niveau de la phrase, il est possible d'identifier, sur la base d'informations syntaxiques et sans recourir à l'identification des liens de reprises, des segments de texte qui expriment une information secondaire : les segments entre parenthèses, les relatives non restrictives, les appositions sont très souvent, sinon toujours, des segments de ce type [19]. Nous serons amenés à identifier ce que nous appelons des « insertions » dans une phrase, les insertions pouvant être vues comme des unités de discours qui sont des satellites de la phrase qui les contient (voir la section 7.3.9, p. 250). Étant donné ces insertions, nous exclurons qu'une expression pronominale apparaissant en dehors d'une insertion renvoie à une expression figurant dans une insertion, ce qui est analogue à l'idée de Cristea et al. que les expressions anaphoriques renvoient de préférence à des expressions figurant dans des unités noyau (voir l'expression de cette contrainte au chapitre 10, section 10.3.3).

6.5 Systèmes d'interprétation automatique des expressions pronominales

Les développements de la linguistique informatique ont conduit au développement dans les années 90 de systèmes de résolution automatique des pronoms effectivement implantés en machine et capables de fonctionner sur des textes quelconques. Notre objectif étant d'implanter un tel système pour le français, nous présentons ici quelques uns de ces systèmes.

Dans la mesure où notre travail a consisté à implanter un système d'hypothèses dans le formalisme d'un outil existant à XRCE (l'outil XIP), et non à définir une architecture et un programme informatique dédié à la résolution des expressions pronominales, nous nous concentrerons plus sur les sources d'informations linguistiques utilisées dans les différents algorithmes présentés ici que sur

[19] Marcu et al. [57], dans leur travail d'annotation des structures rhétoriques, prennent en compte des unités de discours de ce type (plus petites que celles de la figure 6.1) et utilisent une relation « apposition ».

la manière dont elles sont utilisées. Par ailleurs, nous intéressent en premier lieu les systèmes qui ont été effectivement évalués, l'évaluation étant le seul moyen de juger de la pertinence des informations utilisées [20].

Ces limites étant posées, on présente en premier lieu ce qui est à notre connaissance le seul système construit pour le français : celui de A. Popescu-Belis [72]. Les autres algorithmes ont tous été définis pour l'anglais [21]. Un seul n'a pas donné lieu à une implantation par son auteur, celui de Hobbs [45], que nous évoquons néanmoins car il constitue traditionnellement une référence au regard de laquelle les approches plus récentes sont comparées. Sont présentés ensuite en détail les algorithmes de Lappin & Leass [55], Kennedy & Boguraev [53], Baldwin [6, 7], Mitkov [60] et Ge et al. [35]. Nous nous concentrons sur ces systèmes car ils sont les mieux documentés et les plus complètement évalués. On termine par une évocation plus succincte de quelques autres systèmes similaires aux précédents en ce qui concerne l'information utilisée dans le processus de résolution.

6.5.1 Popescu-Belis : résolution de la référence en français

À notre connaissance, un seul système de résolution des pronoms dans des textes libres existe pour le français [22]. Il a été développé par A. Popescu-Belis dans le contexte plus large d'un système de « résolution de la référence » (voir, principalement, [72, chapitre VII]). Le système de Popescu-Belis vise à traiter l'ensemble des phénomènes de coréférence. Pour ce faire il manipule des « représentations mentales », objets conceptuels représentant les êtres dénotés par les expressions. Le texte est analysé linéairement de gauche à droite et pour chaque « expression référentielle » rencontrée, le système détermine si celle-ci doit être rattachée à une représentation mentale existante (il y a reprise avec coréférence) ou donner lieu à la création d'une nouvelle représentation mentale.

Popescu-Belis donne peu de détails sur les règles ou facteurs qui déterminent le rattachement ou non d'une expression à telle ou telle représentation men-

[20] Ces choix nous ont conduits à mettre de côté bon nombre de travaux, en particulier :

– Les premiers systèmes effectivement implantés, qui s'appliquent à des domaines et langages très restreints, par exemple le système SHRDLU de T. Winograd [95], restreint à un univers de blocs, le système de F. Günthner et H. Lehmann [42], restreint à l'interrogation d'une base de données, le système de D. Carter [17], restreint à des histoires courtes rédigées dans une syntaxe simple et avec un lexique réduit.

– Les propositions qui se concentrent plus sur une architecture que sur la définition effective de règles de résolution, par exemple les propositions de J. Carbonnel et R. Brown [16], C. Aone et D. McKee [4], ou encore D. Byron et J. Tetreault [15].

– Bon nombre de systèmes que nous jugeons insuffisamment documentés, en l'occurrence, les systèmes utilisés par les participants aux conferences MUC (*Message Understanding Conferences*) [38]. Comme, dans ce cadre, la tâche d'interprétation des pronoms est loin d'être la seule à accomplir par le système, la description du module correspondant est souvent très réduite.

[21] L'approche de Mitkov a aussi été appliquée au polonais et à l'arabe.

[22] Si on fait abstraction des systèmes fonctionnant dans des contextes très restreints, tel que, par exemple, celui de M. Rolbert [79], limité à l'interface d'interrogation d'une base de données.

tale (voir [72, p. 210]). Trois règles sont en fait utilisées : l'accord en genre et en nombre entre deux expressions coréférentes et une règle d'accord « sémantique » entre les expressions coréférentes, celle-ci s'appliquant pour les syntagmes non pronominaux. Outre ces règles de base, Popescu-Belis fait usage de facteurs d'« activation » dans le style des facteurs de saillance de Lappin & Leass [55] (voir ci-dessous).

D'un point de vue technique, le programme de Popescu-Belis nécessite une intervention manuelle au niveau de l'analyse syntaxique (correction ou sélection d'une analyse produite par un analyseur LFG) et utilise une ressource de sémantique lexicale restreinte aux termes du corpus.

Dans ses publications les plus récentes, Popescu-Belis ne donne pas de résultats sur la tâche spécifique de l'interprétation des pronoms. Dans un article de 1997 [73], sont indiqués les résultats d'une « expérience » où 62 % des pronoms sont correctement rattachés.

6.5.2 Hobbs : un « algorithme naïf » utilisant seulement l'information morpho-syntaxique

Dès 1976, Jerry Hobbs [45] proposait un « algorithme naïf » de résolution des pronoms qui fait usage de la seule information donnée par l'analyse syntaxique du texte. Cet algorithme fait date dans la mesure où les systèmes implantés jusqu'à vingt ans plus tard lui sont toujours comparés.

Étant donné un pronom dans une phrase p_i, l'algorithme de Hobbs est formulé comme une procédure de parcours de l'arbre syntaxique de la phrase p_i et éventuellement de la phrase précédente, en partant du nœud NP (syntagme nominal) qui domine immédiatement le pronom en question, en remontant dans l'arbre, puis en parcourant celui-ci de gauche à droite [23] et en largeur d'abord jusqu'à ce qu'un nœud NP qui satisfasse les contraintes d'accord en genre et nombre avec le pronom soit rencontré, auquel cas ce nœud est proposé comme antécédent. Cet algorithme, dont nous nous contentons de donner ici une vue très approximative est détaillé, non seulement dans [45], mais aussi dans quelques articles d'auteurs différents, en particulier [55], [6, p. 113] et [92].

Walker [92] remarque que l'algorithme de Hobbs exprime une préférence générale pour un antécédent figurant dans la même phrase que le pronom et qui soit le plus proche du pronom. Il exprime également une préférence pour les antécédents sujets puisque le sujet est souvent le premier nœud NP rencontré lorsqu'on parcourt l'arbre syntaxique à partir du nœud racine de gauche à droite et en largeur. Viennent ensuite les syntagmes nominaux objets directs et indirects. Les syntagmes prépositionnels compléments de syntagmes nominaux sont des candidats moins probables puisqu'ils sont enchâssés plus profondément dans l'arbre.

[23] Sauf dans une étape particulière qui gère les cas de cataphore par un parcours de l'arbre sur la droite du pronom.

L'algorithme de Hobbs n'a pas été implanté ; il a été évalué manuellement avec une analyse syntaxique supposée parfaitement correcte. Sont traitées les formes des pronoms de troisième personne [24] à l'exclusion des occurrences du pronom *it* impersonnel ou renvoyant à une proposition. Le taux de succès obtenu est de 88,3 %, un score que peu d'approches récentes atteignent.

6.5.3 Lappin & Leass

En 1994, S. Lappin et H. Leass [55] proposent un algorithme de résolution des pronoms pour l'anglais, appelé RAP (« Resolution of Anaphora Procedure »). Cet algorithme mérite qu'on s'y arrête dans la mesure où les autres systèmes que nous présenterons et notre propre système font usage d'une procédure et/ou d'une information similaire. Nous nous autorisons quelques simplifications, par rapport à [55], dans notre présentation de RAP.

Les principaux composants du système RAP sont les suivants :

- deux filtres excluant, sur la base d'informations syntaxiques et morphologiques, respectivement, la coréférence entre un pronom et un syntagme nominal de la même phrase ;
- une procédure pour identifier les pronoms « pléonastiques (sémantiquement vides) » ;
- un algorithme de liage identifiant l'antécédent d'un pronom réfléchi ou réciproque dans la même phrase que le pronom ;
- une procédure d'assignation d'une valeur de « saillance » pour chaque syntagme nominal ou classe de syntagmes nominaux coréférents entre eux.

La valeur de saillance pour un syntagme nominal SN_i donné est calculée en fonction des valeurs présentées dans le tableau 6.2. Si le syntagme SN_i satisfait une (ou plusieurs) condition(s) exprimée(s) dans la colonne de gauche, la valeur correspondante (ou la somme des valeurs correspondantes) lui est assignée comme valeur de saillance. Les deux dernières valeurs du tableau sont pertinentes pour un couple (P_i, SN_i) où SN_i est un candidat antécédent pour le pronom P_i.

La valeur de saillance pour un syntagme donné est également déterminée, de manière cumulative, en fonction des éventuelles autres expressions avec lesquelles ledit syntagme est coréférent, si bien que cette valeur est en fait associée à des « référents de discours ». Enfin, la valeur de saillance d'un référent donné est diminuée progressivement à mesure que l'analyse du texte se poursuit (de gauche à droite), si aucune expression ne renvoie à ce référent.

[24] En règle générale, les formes correspondant à ce que nous appelons « déterminants possessifs » en français (*his, her, its, their*) sont appelées « pronoms » par les auteurs anglophones. Ceux-ci sont donc traités par l'algorithme de Hobbs.

Condition	V
SN_i est dans la phrase courante	+100
SN_i est sujet	+80
SN_i apparaît dans une construction existentielle (ex. *There are* SN_i)	+70
SN_i est complément d'objet direct	+50
SN_i est complément d'objet indirect ou oblique	+40
SN_i n'est pas enchâssé dans un autre syntagme nominal	+80
SN_i n'est pas enchâssé dans un syntagme prépositionnel adverbial	+50
SN_i suit P_i (cataphore)	−175
SN_i et P_i occupent la même fonction (parallélisme des fonctions)	+35

TAB. 6.2 – Facteurs de saillance utilisés par Lappin & Leass.

Cela étant, la procédure d'identification d'un antécédent pour un pronom donné est la suivante [25] :

1. créer une liste d'antécédents possibles, où les antécédents possibles sont vus comme les référents les plus récemment évoqués ;
2. calculer la valeur de saillance pour chaque antécédent ;
3. éliminer les candidats qui ne satisfont pas les conditions posées par les filtres morphologique et syntaxique ;
4. sélectionner l'antécédent qui a la plus forte valeur de saillance ;
5. si plusieurs antécédents restent possibles, sélectionner le plus proche.

On retrouve dans les facteurs de saillance de Lappin & Leass des préférences qui sont exprimées indirectement dans l'algorithme de Hobbs : la préférence pour un antécédent figurant dans la même phrase, la préférence pour un antécédent sujet, la moins forte probabilité pour un antécédent figurant dans un syntagme prépositionnel complément d'un syntagme nominal.

Notre propre système est en partie comparable à celui de Lappin & Leass dans la procédure qui consiste à sélectionner dans un premier temps un ensemble d'antécédents possibles, éliminer ceux qui ne satisfont pas un certain nombre ce contraintes morphologiques et syntaxiques, puis sélectionner un antécédent parmi les antécédents possibles restant. Cependant, notre approche de cette dernière étape sera sensiblement différente.

Contrairement à l'algorithme de Hobbs, l'algorithme de Lappin & Leass a été implanté en machine. Il a été évalué sur un corpus de manuel technique avec une analyse syntaxique corrigée manuellement. 86 % des pronoms du corpus ont

[25] La procédure présentée est celle qui s'applique aux pronoms non réfléchis et non réciproques, dans une version simplifiée.

été correctement interprétés par RAP. Lappin & Leass ont implanté et évalué l'algorithme de Hobbs sur le même corpus que celui utilisé pour l'évaluation de RAP ; le taux de succès de l'algorithme de Hobbs est de 82 % sur ce corpus.

6.5.4 Kennedy & Boguraev : une approche « sans analyseur »

C. Kennedy et B. Boguraev [53] proposent une version modifiée et étendu de l'algorithme de Lappin & Leass, qui ne nécessite pas une analyse syntaxique complète et en profondeur du texte en entrée [26]. Les auteurs ajoutent deux facteurs de saillance à ceux définis dans [55] : une valeur de 65 est associée à un syntagme nominal possessif et une valeur de 50 à un syntagme figurant dans le contexte courant, un tel contexte étant un « segment de texte cohérent dans son sujet » [27].

Le système de Kennedy & Boguraev obtient un taux de succès de 75 %, sensiblement inférieur au score obtenu par l'algorithme de Lappin & Leass. Cela s'explique probablement par le fait que l'information en entrée du système est moins riche et est obtenue de manière entièrement automatique. Les auteurs avancent également l'hypothèse que leur corpus d'évaluation (articles de presse de genres variés) est plus complexe à analyser que le corpus de manuels techniques utilisé dans [55].

Signalons que R. Stuckardt [85] a défini un système qui utilise des préférences similaires à celles qui sont utilisées par Lappin & Leass. L'évaluation du système dans un contexte comparable à celui du système de Kennedy et Boguraev (analyse syntaxique potentiellement incomplète ou erronée) donne des résultats comparables à ceux qui sont obtenus par ces derniers.

6.5.5 Baldwin : le système CogNIAC

B. Baldwin [6, 7] a conçu un programme de résolution de l'anaphore en général, dans lequel seules des règles décrivant l'interprétation des pronoms ont été implantées [28]. Le système CogNIAC se distingue des autres systèmes présentés ici en ce qu'il ne vise pas à fournir un antécédent unique pour chaque pronom. L'objectif de Baldwin est de distinguer les pronoms qui peuvent être résolus avec une grande précision, étant donné les connaissances dont dispose le système, et ceux dont l'interprétation nécessiterait une source de connaissance additionnelle. Une réponse n'est donné par le système que pour les pronoms du premier type.

[26] D'où l'appellation d'approche « sans analyseur », qui est un peu abusive dans la mesure où ces auteurs utilisent un analyseur qui donne la fonction syntaxique de chaque unité lexicale du texte en entrée.

[27] Peu de détails sont donnés pour ces deux facteurs.

[28] Il semble que le système ait été étendu par la suite, mais nous nous limitons ici au traitement des pronoms.

L'analyse préliminaire du texte consiste en un étiquetage morpho-syntaxique, la reconnaissance des syntagmes nominaux noyau et la détection des propositions. Étant donné un pronom P_i et un ensemble d'antécédents possibles pour ce pronom (les antécédents possibles étant des représentations des entité de l'univers de dénotation associé au texte qui précède et qui sont compatibles avec le genre, le nombre et les restrictions de coréférence associés au pronom), les règles utilisées pour l'interprétation de P_i sont les suivantes :

1. S'il existe un seul antécédent possible dans le discours, sélectionner cet antécédent ;
2. Si P_i est un pronom réfléchi, sélectionner l'antécédent le plus proche ;
3. S'il existe un seul antécédent possible dans la phrase précédente et dans la phrase courante, sélectionner cet antécédent ;
4. Si P_i est un pronom possessif déterminant un syntagme SN_i et s'il existe dans la phrase précédente un syntagme nominal SN_j identique à SN_i et déterminé par un possessif P_j, sélectionner comme antécédent l'entité associée à P_j ;
5. S'il existe un seul antécédent possible dans la phrase courante, sélectionner cet antécédent ;
6. Si P_i est le sujet de la phrase courante et le sujet de la phrase précédente ne contient qu'un antécédent possible, sélectionner cet antécédent ;
7. Dans tous les autres cas, P_i est laissé sans interprétation.

Baldwin [7] donne une évaluation de ce système dans un contexte assez restreint : le système traite les pronoms singuliers dénotant des êtres humains seulement (*he, she, him*...) sur des textes narratifs mettant en jeu deux personnages de même sexe (cela pour maximiser l'ambiguïté de résolution des pronoms). Le rappel (nombre de pronoms correctement résolus sur l'ensemble des pronoms) est de 64 % et la précision (nombre de pronoms correctement résolus sur l'ensemble des pronoms pour lesquels le système donne une interprétation) est de 92 %.

Pour permettre une comparaison avec les approches qui proposent un antécédent pour chaque pronom (p. ex. l'algorithme de Hobbs ou RAP, voir ci-dessus), Baldwin complète son système par deux règles, que nous numéroterons 7a et 7b :

7a S'il existe dans la proposition courante un antécédent possible qui est coréférent avec une expression de la proposition ou phrase précédente, sélectionner cet antécédent [29] ;

7b Sélectionner l'antécédent le plus récent dans le texte.

Avec ces deux règles ajoutées aux précédentes, le système de Baldwin obtient un taux de succès de 77,9 %.

[29] Cette règle fait appel aux notions définies dans la théorie du centrage. Nous en donnons une formulation simplifiée. Si plusieurs antécédents satisfont la condition, celui qui est coréférent avec le sujet est préféré.

D'après l'évaluation effectuée par Baldwin, la règle 3 est celle qui s'applique le plus souvent (35 % des pronoms du corpus sont traités par elle). On remarquera que la règle 5 est un cas particulier de la règle 3 et que la règle 4, qui s'applique très rarement (sur 1 % des pronoms du corpus), a valeur d'exception à cette règle. On peut faire l'hypothèse que la valeur des règles 3 et 5 vient du fait que les reprises internes à la phrase sont en général nettement plus fréquentes que les reprises inter-phrases.

Baldwin [7] propose également une évaluation d'une version étendue de CogNIAC : le système traite le pronom *it* en sus des pronoms précédemment traités, les règles 4, 7a et 7b sont supprimées parce que jugées inappropriées pour le domaine considéré, en l'occurrence des articles de presse utilisés pour la campagne MUC-6 [38], d'autres sont ajoutées (voir [7] pour une description de ces règles). Les résultats obtenus, avec un processus d'analyse entièrement automatique, sont de 75 % pour le rappel et 73 % pour la précision (dans cette version, le système donne une réponse pour chaque pronom).

Baldwin prend soin de distinguer son système de l'algorithme de Hobbs ou du système de Lappin & Leass en insistant sur le fait qu'il ne s'engage pas à fournir une réponse pour tout pronom. Cette caractéristique n'est cependant pas si spécifique, dans la mesure où un système tel que celui de Lappin & Leass pourrait aisément être adapté pour ne rattacher effectivement un pronom à un antécédent que si celui-ci a une valeur de saillance nettement supérieure à celle des autres antécédents possibles.

En revanche, le fait que Baldwin sélectionne l'antécédent d'un pronom sur un critère absolu est plus particulier à son système. Chaque règle est susceptible soit de n'avoir aucun effet sur l'ensemble des antécédents possibles, soit de sélectionner un et un seul antécédent parmi l'ensemble des antécédents possibles. Une telle approche a l'avantage de permettre une évaluation moins diffuse des règles utilisées, le degré de validité de chacune pouvant être mesuré (par exemple la règle 1 a été évaluée comme produisant toujours une réponse correcte dans les deux évaluations, la règle 3 comme ayant une précision de 96 % sur un corpus et de 72 % sur un autre corpus).

6.5.6 Mitkov : une approche « robuste et pauvre en connaissance »

R. Mitkov [60] conduit depuis plusieurs années des travaux sur la résolution des anaphores pronominales qui l'ont mené à expérimenter différentes approches. Nous nous concentrerons ici sur la plus récente, dite « approche robuste et pauvre en connaissance » [30].

[30] Pour une discussion des travaux plus anciens de Mitkov, nous renvoyons à son propre état de l'art [61].

L'approche robuste et pauvre en connaissance de Mitkov comprend les étapes suivantes, étant donné un pronom P_i :

1. étiquetage morpho-syntaxique désambiguïsé du texte en entrée ;
2. identification des syntagmes nominaux qui précèdent P_i dans un espace de deux phrases ;
3. vérification de l'accord en genre et nombre ;
4. assignation d'un ensemble d'« indicateurs d'antécédent », décrits dans le tableau 6.3 ;
5. sélection de l'antécédent ayant obtenu le meilleur score.

Cette approche est similaire à celle de Lappin & Leass en ce qui concerne les « indicateurs d'antécédent », qui fonctionnent sur le même principe que les valeurs de saillance de Lappin & Leass, à cette différence près qu'ils ne s'appliquent que pour un syntagme nominal donné et non pour une classe de syntagmes associés à un référent. Notons cependant que l'information syntaxique utilisée est nettement moins riche que celle qui est utilisée par Lappin & Leass, puisqu'elle se limite à une identification des syntagmes nominaux.

On retrouve dans les indicateurs de Mitkov (voir le tableau 6.3 quelques préférences que nous avons rencontrées explicitement ou implicitement précédemment : la préférence pour les syntagmes nominaux non prépositionnels, la préférence pour un antécédent figurant dans la même phrase que le pronom et, de manière indirecte dans la préférence pour le thème, une préférence pour ce qui est sans doute le plus souvent le sujet de la phrase (Mitkov ne dispose pas de l'information sur les fonctions des syntagmes nominaux). En ce qui concerne cette dernière préférence, Mitkov signale que son corpus de manuels techniques contient beaucoup de phrases impératives, d'où peut-être la moins grande importance accordée au sujet dans ce système.

L'approche de Mitkov a été implantée pour l'anglais, le polonais et l'arabe (avec quelques indicateurs supplémentaires dans ce dernier cas) [63]. Évaluée sur un corpus de manuels techniques, elle donne un taux de succès de l'ordre de 90 % pour chacune des trois langues. Mitkov signale cependant que les résultats sont susceptibles de varier d'un genre à l'autre. Par ailleurs, une autre évaluation, réalisée par l'auteur lui-même [8] et visant à comparer trois approches (l'approche de Mitkov, celle de Baldwin (voir ci-dessous) et celle de Kennedy et Boguraev), donne des résultats nettement inférieurs à ce qui avait été obtenu dans un premier temps pour chacune des trois approches (56,9 % pour l'approche de Mitkov). Nous reviendrons sur les problèmes d'évaluation au chapitre 12.

Mitkov [62, 66] propose également une évaluation du pouvoir de décision et de l'« indispensabilité » (ou « importance relative ») de chaque indicateur I.

Le pouvoir de décision d'un indicateur I est défini comme le nombre de fois où un candidat ayant reçu l'indicateur I a été sélectionné sur le nombre total de

Défini/indéfini	
SN_i est un syntagme nominal indéfini	-1

Thème	
SN_i est le premier SN d'une phrase non impérative	$+1$

Verbes indicateurs	
SN_i est le premier SN après un verbe tel que *discuss, present*	$+1$

Réitération lexicale	
SN_i est répété au moins deux fois dans le même paragraphe	$+2$
SN_i est répété une fois dans le même paragraphe	$+1$

Préférence pour le titre	
SN_i apparaît dans le titre de la section à laquelle appartient la phrase	$+1$

Syntagmes nominaux « non prépositionnels »	
SN_i est enchâssé dans un syntagme prépositionnel	-1

Patron de collocation	
SN_i et P_i sont sujets ou objets de deux occurrences d'un même verbe	$+2$

Référence immédiate	
SN_i et P_i sont objets respectivement de V_i et V_j dans deux propositions coordonnées ou liées par une conjonction telle que *before, after*	$+2$

Distance référentielle	
SN_i précède P_i dans la même phrase	$+2$
SN_i est dans la phrase qui précède celle où figure P_i	$+1$
SN_i est dans la 2e phrase qui précède celle où figure P_i	0
SN_i est dans la 3e phrase qui précède celle où figure P_i	-1

Préférence pour les termes	
SN_i est un terme du domaine dont parle le texte	$+1$

TAB. 6.3 – Indicateurs d'antécédent selon Mitkov [60].

fois où I a été assigné à un syntagme nominal. Les indicateurs ayant le plus fort pouvoir de décision sont la référence immédiate, la préférence pour les syntagmes non prépositionnels et la préférence pour le patron de collocation. Les indicateurs ayant le plus faible pouvoir de décision sont la distance référentielle, la préférence pour les termes et la préférence pour le thème.

L'indispensabilité d'un indicateur I est donné par la formule $S - S_{-I}/S$ où S est le taux de succès obtenu part le système complet et S_{-I} est le taux de succès obtenu par le système sans l'indicateur I. Les indicateurs les plus indispensables sont l'indicateur de distance référentielle, la préférence pour le thème (en l'occurrence le premier syntagme nominal d'une phrase) et la préférence pour les syntagmes nominaux non prépositionnels.

6.5.7 Ge et al.

Ge et al. [35] ont implanté un système qui, étant donné un corpus (le Penn Tree Bank) annoté de manière approprié calcule la probabilité d'une coréférence pronom-antécédent, étant donné des facteurs similaires à ceux qui sont utilisés en général (distance, accord, patrons de cooccurrence, répétition des mentions, etc.). Appliqué à un échantillon du même corpus n'ayant pas servi à l'entraînement, le système acquis interprète correctement 82,5 % des pronoms.

L'évaluation de l'apport des différentes sources d'information indique l'importance de la syntaxe : un système consistant à retenir comme antécédent l'expression la plus proche produirait un taux de succès de 43 %, l'utilisation d'information sur la structure syntaxique (sous la forme d'une version modifiée de l'algorithme de Hobbs) produit un taux de succès de 65,3 %. L'ajout des contraintes d'accord porte ce chiffre à 75,7 %, d'où les auteurs concluent à l'importance de cette information. L'utilisation d'une information comparable aux patrons de cooccurrence de Dagan et Itai ne produit qu'une amélioration de 2,2 % (77,9 %), un résultat comparable à celui obtenu par l'ajout de ce type d'information à l'algorithme de Lappin & Leass (voir p. 195). Enfin, la prise en compte du nombre de fois où un référent est évoqué produit une amélioration de 4,6 %.

L'intérêt du système de Ge et al. est que les poids utilisés sont estimés selon une méthode objective (mesures statistiques à partir d'un corpus), alors que les poids utilisés par Lappin & Leass ou Mitkov sont déterminés au départ de manière intuitive, puis raffinés par tests successifs. Cette approche permet d'envisager aisément l'adaptation du système à un corpus ou type de corpus donné : la procédure d'acquisition des poids serait la même pour chaque corpus mais résulterait éventuellement en des poids différents.

6.5.8 Autres systèmes

Pour terminer, nous évoquons brièvement quelques autres systèmes de résolution automatique des pronoms, sans viser à l'exhaustivité tant les systèmes

existants sont comparables. On y retrouve en effet le même type d'information que celle qui est utilisée par les systèmes décrits précédemment. Les systèmes de Connolly et al. et de Ge et al. se distinguent cependant des autres en ce qu'ils mettent en jeu des approches d'apprentissage automatique.

Nasukawa

T. Nasukawa [65] obtient des résultats similaires à ceux obtenus par Mitkov (93,8 %, évaluation sur 84 pronoms), sur le même type de corpus (manuels techniques). L'information utilisée est la suivante [31] :

- patrons de collocation apparaissant dans le texte précédant la phrase où apparaît le pronom. Les patrons de collocations sont des données du même type que celles qu'utilisent Dagan et Itai (voir page 193). Nasukawa prend en compte les patrons de collocation mettant un jeu un synonyme du candidat antécédent.
- fréquence de répétition d'un même lemme dans le texte précédant la phrase où apparaît le pronom.
- proximité et position de l'antécédent potentiel dans la phrase. Préférence est donnée aux expressions les plus à gauche dans la phrase.

Nasukawa analyse l'apport de ces trois sources d'information. L'utilisation de la seule information de proximité et de position produirait un résultat de 82,5 % et celle de la seule répétition des termes dans le texte un résultat de 60,7 %. Les patrons de collocation n'interviennent que dans 26,2 % des cas, mais n'engendrent aucune erreur. Dans la moitié des cas où les patrons de collocation interviennent, les autres sources d'information auraient conduit à la sélection d'un antécédent incorrect.

GPLSI

Le GPLSI de l'université d'Alicante mène une recherche très active dans la résolution d'anaphore. Un système a été implanté pour l'espagnol, susceptible de fonctionner sur des textes écrits [32] ou des dialogues [58], ou dans le contexte d'un système de traduction automatique [68]. La procédure générale de résolution est du même type que celle utilisée par Mitkov ou Lappin & Leass : un premier module détermine l'espace dans lequel figure l'antécédent et retourne une liste d'antécédents possibles, un deuxième module filtre la liste en éliminant les antécédents possibles qui ne satisfont pas les conditions d'accord et restrictions syntaxiques et un dernier module applique des préférences. Ces dernières sont peu

[31] Étant donné un ensemble d'antécédents potentiels extrait dans les deux phrases précédant celle qui contient le pronom, puis réduit après une première phase de filtrage en fonction de l'accord et de contraintes syntaxiques.

documentées dans les articles que nous avons pu consulter. Si le système a bien été évalué, il l'a été sur un ensemble réduit de pronoms (80 pronoms dans [58], un corpus de 9600 mots dans [32]).

Connolly et al.

Connolly et al. [22] décomposent le problème de sélection d'un référent pour une expression anaphorique en sous-problèmes consistant à catégoriser successivement des paires de candidats pour une expression anaphorique donnée. La catégorisation en question est une catégorisation en deux classes consistant simplement à dire lequel des deux candidats est le meilleur.

L'information utilisée par Connolly et al. est la suivante [32] :

- fonction grammaticale de l'expression anaphorique et des candidats : sujet, objet direct, objet indirect, autres fonctions ;
- distance de l'expression anaphorique au candidat : même phrase, phrase précédente, deuxième ou troisième phrase précédente, autre distance ;
- relation de précédence entre deux candidats ;
- accord en nombre entre l'expression anaphorique et le candidat ;
- accord en genre entre l'expression anaphorique et le candidat.

Étant donné un corpus d'articles de presse où les liens de coréférence ont été spécifié manuellement et les information indiquées ci-dessus ont été associées aux expressions concernées par des moyens automatiques (donc susceptible de produire des erreurs), les auteurs construisent automatiquement différents catégoriseurs, dont les résultats sont décrits dans [22]. Le meilleur catégoriseur obtient un taux de succès de 55,3 %, un score assez nettement inférieur à celui obtenu par les algorithmes décrits plus haut. Les auteurs ne donne pas d'explication à cette différence ; elle pourrait être due aux défauts du système produisant l'analyse du texte en entrée (aussi bien pour les données d'entraînement que pour les données d'évaluation) et/ou à la relative pauvreté des informations utilisées.

6.6 Mise en perspective de notre propre système

Le système d'interprétation des expressions pronominales que nous avons implanté est assez proche de ceux de Lappin & Leass, Mitkov ou Baldwin en ce qui concerne l'information utilisée et une partie de la stratégie de résolution.

L'information que nous utiliserons sera essentiellement la structure syntaxique du texte, les informations de nature morphologique (genre et nombre), une quantité réduite d'information de sémantique lexicale (noms décrivant des personnes,

[32] Le système de Connolly et al. gère à la fois les expressions pronominales et les syntagmes nominaux définis. On ignore ici l'information qui est utilisée pour les seuls syntagmes nominaux définis.

noms de mesures et de temps) et l'idée que plusieurs pronoms dans un même segment de discours ont de préférence la même dénotation, ce que nous qualifierons d'une préférence pour la « cohésion » du discours. Nous ne disposons pas d'informations sur les restrictions de sélection ou les patrons de cooccurrence de termes, encore moins d'informations qui pourraient nous permettre d'intégrer des inférences du type de celles qui sont présentées dans la section 6.3.

En ce qui concerne la stratégie de résolution des expressions pronominales, notre système, comme ceux de Mitkov et Lappin & Leass, passe par une première étape de sélection d'un ensemble d'antécédents possibles pour chaque pronom, l'application de contraintes morpho-syntaxiques visant à réduire cet ensemble, et la sélection d'un antécédent unique par application de préférences. C'est surtout au niveau des préférences que notre système se distingue : celles-ci ne sont pas implantées comme un mécanisme d'assignation de poids, mais comme un ensemble de formules ordonnées qui visent à réduire progressivement l'ensemble des antécédent potentiels à un élément [33]. C'est l'ordre d'application des différentes préférences qui spécifie le poids qu'on accorde à chacune d'entre elles.

Une comparaison plus fine de notre système avec les systèmes existants sera possible après sa présentation détaillée dans la suite de la thèse.

[33] La pertinence de ces préférences ordonnées, par rapport à un système de préférences pondérées, sera discutée au chapitre 11, section 11.1.1.

Chapitre 7

Analyse syntaxique en entrée du système

Les formules décrivant l'interprétation des expressions pronominales que nous avons retenues (voir section 5.1) sont exprimées en fonction de l'information fournie par l'analyseur syntaxique du français développé à XRCE. L'objet du présent chapitre est de décrire la sortie produite par cet analyseur syntaxique.

Le système de règles décrivant la syntaxe du français est implanté dans le formalisme de l'outil XIP, outil qui, étant donné un texte en entrée, génère, à partir des règles formulées, l'analyse dudit texte. Notre système de résolution des pronoms étant lui-même implanté dans le formalisme de XIP, le présent chapitre a également pour objectif de décrire les structures de données que XIP est susceptible de manipuler.

Tel qu'il est implanté à l'heure actuelle, le système de règles décrivant la structure syntaxique des phrases du français vise à fournir une analyse finale en termes de « dépendances », mais il utilise pour cela une analyse préliminaire sous la forme d'un arbre syntaxique partiel [1]. Nous ferons usage à la fois de l'information fournie par l'arbre syntaxique partiel et par les dépendances et présentons donc ici ces deux structures.

La présentation de l'arbre syntaxique fait l'objet de la section 7.1. Les objets qui constituent l'arbre syntaxique sont des nœuds, auxquels sont associés des « traits » qui permettent de les caractériser. La section 7.2 décrit les principes généraux des traits dans XIP et la section 7.3 décrit les différentes catégories de nœuds que ces traits spécifient, catégories qui seront utilisées dans notre système de résolution des pronoms. Enfin, la section 7.4 décrit les dépendances qui rendent compte de la structure syntaxique des phrases en sortie du processus d'analyse syntaxique.

[1] Ce type d'analyse syntaxique est similaire à celui qui était produit par l'analyseur IFSP, développé précédemment à XRCE (voir [1] et [2]). Une présentation succincte du système XIP et de l'analyse effectuée pour le français est disponible dans [3].

Le travail présenté dans cette section est principalement celui de Salah Aït-Mokhtar et Jean-Pierre Chanod pour l'analyse syntaxique, auxquels s'ajoute Claude Roux pour ce qui concerne la définition et l'implantation du système XIP lui-même. Certains éléments d'analyse nécessaires à notre système de résolution des pronoms ont cependant été ajoutés par nous-mêmes. Une section finale (7.5) résume notre apport dans la définition des structures d'information décrites dans ce chapitre.

7.1 Arbre syntaxique

On distingue dans l'arbre syntaxique deux types de nœuds : les nœuds lexicaux et les nœuds non lexicaux. Les nœuds lexicaux sont les nœuds qui dominent immédiatement les unités lexicales, ces unités étant les feuilles de l'arbre ; les nœuds non lexicaux sont ceux qui dominent les nœuds lexicaux. Ces deux types de nœuds sont présentés dans les deux sections suivantes [2].

7.1.1 Nœuds lexicaux

Les nœuds lexicaux sont les nœuds qui dominent immédiatement les unités lexicales. Ces nœuds sont spécifiés à partir des informations fournies par l'analyseur morphologique et le processus de désambiguïsation des unités lexicales (voir la figure 5.1 page 180 et les figures 5.3 et 5.4 page 182).

À partir de l'analyse de la figure 5.4, l'analyseur construit une sorte de proto-arbre syntaxique, tel que chaque unité lexicale a pour parent un nœud dont l'étiquette est celle de la partie du discours associée à l'unité lexicale. Ce proto-arbre a la forme suivante :

```
  .     .     .     .     .
  .     .     .     .     .
  |     |     |     |     |
PRON  VERB  DET  NOUN  SENT
  |     |     |     |     |
  Il  ferme  la  marche  .
```

Les principales étiquettes possibles pour les nœuds lexicaux sont les suivantes :

`NOUN`, pour les noms communs et noms propres ;
`PRON`, pour les pronoms ;
`ADJ`, pour les adjectifs ;
`VERB`, pour les formes verbales, participes inclus ;

[2] La figure B.1 (annexe B), donne l'arbre syntaxique complet construit par l'analyseur syntaxique du français pour un court extrait d'un article de *La Tribune*.

`ADV`, pour les adverbes ;
`PREP`, pour les prépositions ;
`COORD`, pour les conjonctions de coordination (p. ex. *et, ou*) ;
`CONJ`, pour les conjonctions autres que les conjonctions de coordination (p. ex. *que, quand, si*) ;
`DET`, pour les déterminants ;
`NUM`, pour les numéraux cardinaux, aussi bien en chiffres (p. ex. *2*) qu'en lettres (p. ex. *deux*) ;
`PUNCT`, signe de ponctuation ne marquant pas une fin de phrase (p. ex. « , », « (», « : ») ;
`SENT`, signe de ponctuation marquant une fin de phrase.

La sémantique de certaines de ces étiquettes — les plus pertinentes pour la tâche que nous nous sommes fixé — sera détaillée plus loin (section 7.3).

7.1.2 Syntagmes et propositions noyau

Nous avons qualifié de « partiel » l'arbre syntaxique construit par l'analyseur du français parce qu'il ne vise pas à décrire l'ensemble de la structure de la phrase, contrairement à ce qui est fait avec les grammaires classiques qui visent à une analyse en constituants immédiats.

L'analyse proposée est une analyse en syntagmes *noyau* et propositions finies *noyau*. Les syntagmes et propositions finies noyau sont définis à partir des notions de syntagmes et propositions de l'analyse en constituants traditionnelle. Un syntagme noyau est la partie d'un syntagme qui va de son début jusqu'à son noyau inclus ; une proposition finie noyau est la partie d'une proposition ayant pour noyau un verbe fléchi qui va de son début jusqu'à son noyau, en l'occurrence ledit verbe fléchi.

La notion de noyau correspond à ce qu'on appelle souvent en linguistique la « tête » d'un syntagme. Le terme « noyau » est emprunté à Grevisse [37, §270] :

> La **subordination** est la relation qui unit, à l'intérieur de la phrase, des éléments qui ne sont pas de même niveau, qui ont des fonctions différentes, dont l'un dépend de l'autre. Ils forment un groupe, un *syntagme*, dans lequel il y a un élément syntaxiquement plus important, le *noyau*, qui est comme le *support* des éléments dépendants, subordonnés, appelés généralement **compléments**.

Soit la phrase,

(1) André Lévy-Lang réunit demain les actionnaires de la compagnie de la rue d'Antin, qui devront entériner les comptes de l'exercice 1997.

Parmi les syntagmes nominaux de cette phrase, on relève les syntagmes suivants :

– les actionnaires de la compagnie de la rue d'Antin

- la compagnie de la rue d'Antin
- la rue d'Antin
- Antin

À ces quatre syntagmes nominaux complets correspondent respectivement les quatre syntagmes nominaux noyau suivants :

- les actionnaires
- la compagnie
- la rue
- Antin

La phrase ci-dessus contient par ailleurs deux propositions finies : la proposition principale *André Lévy-Lang réunit demain les actionnaires de la compagnie de la rue d'Antin* et la proposition relative *qui devront entériner les comptes de l'exercice 1997*. À ces deux propositions complètes correspondent les deux propositions finies noyau suivantes :

- André Lévy-Lang réunit
- qui devront

Une proposition finie noyau n'est identifiée par l'analyseur qu'à la condition qu'elle contienne un syntagme verbal noyau avec verbe fléchi, ce syntagme étant le noyau de la proposition. On emploiera désormais le terme « proposition noyau », ou simplement « proposition » pour faire référence à une proposition finie noyau.

L'arbre syntaxique construit par le système de règles défini pour le français se limite donc à un découpage en syntagmes et propositions noyau. Pour le syntagme nominal *le président de la société*, par exemple, une grammaire syntagmatique classique proposerait l'arbre syntaxique de la figure 7.1, avec les étiquettes `NP` et `PP` pour les syntagmes nominaux et les syntagmes prépositionnels, respectivement. L'analyse effectuée pour le français en utilisant le système XIP pour ce même syntagme sera plus « plate ». Elle est présentée figure 7.2. Dans cet arbre le nœud étiqueté `GROUPE` est la racine construite par défaut par le système XIP pour l'arbre d'analyse d'une séquence d'unités lexicales donnée.

Le premier de ces deux arbres syntaxiques contient une information plus riche : l'enchâssement du syntagme *de la société* dans le syntagme plus large *le président de la société* explicite le rapport de complémentation à l'intérieur du syntagme, alors que dans l'analyse effectuée avec XIP pour le français, les syntagmes noyau *le président* et *de la société* sont deux syntagmes autonomes. L'information sur la complémentation sera exprimée au moyen de « dépendances » (voir la section 7.4). De manière générale, l'objectif premier des auteurs de l'analyseur du français tel qu'il est implanté dans XIP est de fournir une analyse syntaxique entièrement exprimée par des dépendances. Dans cette optique, la construction de l'arbre syntaxique partiel n'est perçue que comme une étape visant à isoler des segments

```
          NP
   -------------
   |          NOUN'
   |      -----------
   |      |         PP
   |      |      --------
   |      |      |      NP
   |      |      |   -------
DET   NOUN   PREP DET  NOUN
 |      |      |   |     |
le  président de  la  société
```

FIG. 7.1 – Exemple d'analyse en constituants

```
          GROUPE
      ---------------
      NP            PP
      |          --------
      |          |      NP
   --------      |   -------
DET   NOUN   PREP DET  NOUN
 |      |      |   |     |
le  président de  la  société
```

FIG. 7.2 – Exemple d'analyse en syntagmes noyau

de la phrase de telle sorte que les dépendances pourront être exprimées dans les règles de manière plus simple. Le choix de cette double structure se justifie par l'impossibilité de détecter, dans le cas général, les frontières des syntagmes complets sur une base purement syntaxique. Cela étant, cet arbre syntaxique existe et nous serons amenés à y faire référence par la suite.

Les étiquettes associées aux nœuds qui correspondent aux différents syntagmes noyau et aux propositions noyau sont les suivantes :

`AP` (« adjectival phrase »), syntagme adjectival noyau ;
`NP` (« noun phrase »), syntagme nominal noyau ;
`PP` (« prepositional phrase »), syntagme prépositionnel noyau ;
`FV` (« finite verb »), syntagme verbal noyau avec forme fléchie ;
`IV` (« infinitive verb »), syntagme verbal noyau avec verbe à l'infinitif ;
`GV` (« gerund verb »), syntagme verbal noyau avec verbe au participe présent ;
`SC` (« sentence chunk »), proposition noyau, c'est-à-dire partie d'une proposition qui va de son début jusqu'au verbe fléchi.

À ces étiquettes de syntagmes et propositions noyau s'ajoutent les deux étiquettes suivantes :

> `INS` (« insertion »), segment de texte entre parenthèses ou entre tirets équivalents à des parenthèses ;
> `BG` (« beginning of group »), nœud dominant une expression ou suite d'expressions susceptibles de marquer le début d'une proposition enchâssée ou coordonnée (par exemple, un pronom relatif, une conjonction).

7.1.3 Nœuds « phrase »

Les nœuds décrits dans les deux sections précédentes visent à décrire la structure syntaxique partielle des phrases. À ces nœuds s'ajoute un nœud qui domine chaque phrase entière. Ce nœud est étiqueté `ST`, pour l'anglais « sentence ».

L'analyseur ne construit pas un arbre syntaxique par phrase, mais un arbre syntaxique unique pour l'ensemble du texte en entrée. La racine de cet arbre est le nœud `GROUPE`, construit par défaut par le système (voir p. 224). Le nœud `GROUPE` domine immédiatement une séquence de nœuds `ST` correspondant chacun à une phrase du texte en entrée. Les nœuds `ST` dominent les différents nœuds décrits dans les deux sections précédentes.

Un nœud `ST` est caractérisé comme dominant immédiatement une séquence de nœuds se terminant à droite par un ou plusieurs nœuds lexicaux dominant l'un des symboles suivants : « `:` », « `;` », « `?` », « `!` » ou « `.` », ou encore « `...` » dans certains contextes ; ces symboles ne peuvent apparaître dans un nœud `ST` qu'en position finale. Les symboles « `?` », « `!` » et « `.` » ne peuvent être dominés que par un nœud `ST` et sont donc toujours considérés comme des fins de phrase. Les deux premiers symboles ne sont considérés comme des fins de phrase que s'ils ne sont pas dominés par un nœud `INS` (« insertion » entre parenthèses).

Étant donné cette caractérisation, un nœud `ST` ne contient pas nécessairement un verbe. Un tel nœud peut éventuellement dominer une séquence limitée à deux nœuds. À titre d'exemple, la phrase suivante, extraite de notre corpus d'étude, est analysée comme contenant deux nœuds `ST` (voir la figure 7.3).

(2) Autre candidat à devoir faire ses preuves : Eureko.

Un autre exemple est présenté sur la figure 7.4, qui donne l'arbre syntaxique construit par XIP pour le texte suivant :

(3) Le chat est fatigué. Il dort.

La présentation sous forme graphique des arbres pouvant requérir un espace horizontal important, nous utiliserons le plus souvent dans la suite de la thèse une notation par parenthésage dont la figure 7.5 donne un exemple. Une suite de la forme `ÉTIQUETTE{...}` correspond à un nœud ayant pour étiquette `ÉTIQUETTE` ;

```
                                        GROUPE
                             --------------------------------
                             |                              |
                            ST                             ST
        -------------------------------------------     -------
        |          |        |       |      |       |     |     |
       NP         IV       ADV     IV     NP      PUNCT  NP  SENT
    --------     ------     |       |   -------    |     |     |
    |      |     |    |     |       |   |     |    |     |     |
   ADJ    NOUN PREP VERB    |     VERB DET  NOUN   |    NOUN   |
    |      |     |    |     |       |   |     |    |     |     |
  Autre candidat à devoir encore faire ses preuves :   Eureko  .
```

Fig. 7.3 – Exemple d'arbre syntaxique (1)

```
                              GROUPE
                        ----------------------
                        |                    |
                       ST                   ST
                 ----------------         --------
                 |        |     |         |      |
                SC       AP   SENT       SC    SENT
              -------     |     |      ------    |
              |     |     |     |      |    |    |
             NP     FV   ADJ    |     NP   FV    |
           ------   |     |     |      |    |    |
           |    |   |     |     |      |    |    |
          DET NOUN VERB   |     |    PRON VERB   |
           |    |   |     |     |      |    |    |
          Le  chat est fatigué  .     Il  dort   .
```

Fig. 7.4 – Exemple d'arbre syntaxique (2)

```
GROUPE{
  ST{SC{NP{DET{Le} NOUN{chat}} FV{VERB{est}}}
     AP{ADJ{fatigué}}
     SENT{.}}
  ST{SC{NP{PRON{Il}} FV{VERB{dort}}}
     SENT{.}}}
```

Fig. 7.5 – Notation par parenthésage pour l'arbre de la figure 7.4

entre les accolades figurent la suite des nœuds ou unités lexicales dominés par le nœud en question. Pour une meilleure lisibilité, certains des niveaux d'enchâssement sont marqués par l'indentation.

Dans la suite de la thèse, nous emploierons le terme « phrase » pour désigner une séquence de nœuds dominés par un nœud `ST`. L'exemple (2) contient donc deux phrases.

7.2 Le système de traits dans XIP

Aux nœuds de l'arbre syntaxique sont associés des « traits ». Un trait vise à exprimer une propriété qu'on associe au nœud concerné, par exemple qu'il appartient à telle ou telle partie du discours, que l'unité lexicale qu'il domine est au singulier ou au pluriel, etc. Un trait prend la forme d'un couple `attribut:valeur`.

À titre d'exemple, les traits associés aux quatre nœuds lexicaux de l'arbre syntaxique construit pour la phrase *Le chat dort.* :

```
GROUPE{ST{SC{NP{DET{Le} NOUN{chat}} FV{VERB{dort}}} SENT{.}}}
```

sont les suivants [3] :

- au nœud `DET` sont associés les traits `det:+`, `def:+`, `masc:+`, `sg:+` (autrement dit, *Le* est un déterminant défini masculin singulier) ;
- au nœud `NOUN` sont associés les traits `noun:+`, `sg:+`, `masc:+` (autrement dit, *chat* est un nom masculin singulier) ;
- au nœud `VERB` sont associés les traits `verb:+`, `ind:+`, `pre:+`, `p3:+`, `sg:+` (autrement dit, *dort* est une forme verbale au présent de l'indicatif, troisième personne du singulier) ;
- au nœud `SENT` est associé le trait `sent:+`.

Nous verrons par la suite que les nœuds lexicaux ne sont pas les seuls auxquels sont associés des traits.

Pour comprendre l'intérêt des traits associés aux nœuds de l'arbre syntaxique, il est nécessaire d'anticiper un peu sur la présentation des fonctionnalités offertes par le système XIP, fonctionnalités qui seront présentées en détail au chapitre 8. L'objet de la présente section est donc de donner une vue générale du système de traits dans XIP et de ce qu'il permet d'exprimer. Les traits effectivement utilisés dans l'analyse du français seront présentés dans la section 7.3.

7.2.1 Déclaration des traits

Les différents traits qui pourront être associés aux nœuds de l'arbre syntaxique sont déclarés par le linguiste dans un fichier spécifique au moyen de formules de

[3] La liste des traits donnée ici pour ces unités n'est pas exhaustive.

la forme :

```
attribut1:{valeur1,valeur2,...,valeurN},
attribut2:{valeur1,valeur2,...,valeurN},
...,
attributN:{valeur1,valeur2,...,valeurN},
```

où `attribut1`, `attribut2`, ..., `attributN` sont des noms d'attributs et `valeur1`, `valeur2`, ..., `valeurN` constituent l'ensemble des valeurs, nécessairement différentes, que peut prendre l'attribut en question.

Un même nœud ne peut avoir deux traits utilisant le même attribut et des valeurs différentes. Par exemple, il est impossible qu'un même nœud ait à la fois le trait `attribut1:valeur1` et le trait `attribut1:valeur2`.

7.2.2 Attributs généraux

Outre les attributs qui entrent dans la composition des traits, le système XIP permet la définition d'« attributs généraux ». Un attribut général est un attribut qui référence un ensemble de traits. Il n'a pas à proprement parler de valeur, mais sera utilisé pour évaluer si un nœud de l'arbre syntaxique a ou non un trait quelconque parmi les traits référencés par l'attribut général.

Exemple. Pour décrire les formes verbales participes, deux traits ont été définis : `partpas:+` et `partpre:+`, pour les participes passés et présents, respectivement. Pour faire référence à une forme verbale qui soit participe passé *ou* participe présent, c'est-à-dire à un participe quelconque, on pourra se donner un attribut général `participe` qui référence l'ensemble constitué des traits ayant pour attribut `partpas` ou `partpre`. Ce qu'on déclare de la manière suivante :

```
participe:[
  partpre:{+},
  partpas:{+}]
```

La nature et l'intérêt des attributs généraux apparaîtra plus clairement avec la présentation des différents types de conditions qui peuvent être posées sur les traits associés à un nœud dans XIP (section 7.2.3).

7.2.3 Conditions sur les traits

Un point central dans une formule (ou règle) XIP est qu'elle fait toujours référence aux nœuds de l'arbre syntaxique. Les traits associés aux nœuds de l'arbre syntaxique vont permettre de spécifier plus ou moins précisément ces références.

On se donne la notation suivante : le symbole « `?` » désigne n'importe quel nœud de n'importe quel arbre syntaxique. Nous dirons que « `?` » dénote l'ensemble

de tous les nœuds possibles et que dans une formule XIP, ce symbole peut être instancié par n'importe quel nœud. Pour spécifier la dénotation de « ? », on peut poser des conditions sur les traits qui lui sont ou ne lui sont pas associés.

Supposons que nous ayons spécifié la déclaration des traits présentée figure 7.6. Dans cette déclaration, les attributs `att1` et `att2` ne peuvent avoir qu'une seule valeur : `+`. L'attribut `att3` peut prendre l'une quelconque des trois valeurs `v1`, `v2` ou `v3`. L'attribut `attG` est un attribut général qui référence les traits `att1:+` et `att2:+`.

```
attG:[
  att1:{+},
  att2:{+} ],

att3:{v1,v2,v3}
```

FIG. 7.6 – Exemple de déclaration de traits

Selon cette déclaration, et étant donné qu'un même nœud ne peut avoir deux traits ayant le même attribut, un nœud peut avoir un ou plusieurs des traits suivants :

`att1:+`
et/ou `att2:+`
et/ou `att3:v1` ou `att3:v2` ou `att3:v3`

On formule des conditions sur la dénotation de « ? » entre crochets à la droite du symbole. Quatre types de conditions sont possibles :

(i) la condition `[att:val]` est satisfaite si le nœud a la valeur `val` pour l'attribut `att` (autrement dit, si le nœud a le trait `att:val`) ;
(ii) la condition `[att:~val]` est satisfaite si le nœud n'a pas la valeur `val` pour l'attribut `att` ;
(iii) la condition `[att]` est satisfaite si le nœud a une valeur, quelle qu'elle soit, pour l'attribut `att`, ou bien, si `att` est un attribut général, si le nœud a au moins un des traits référencés par l'attribut `att` ;
(iv) la condition `[att:~]` est satisfaite si le nœud n'a pas de valeur pour l'attribut `att`, ou bien, si `att` est un attribut général, si le nœud n'a aucun des traits référencés par l'attribut `att`.

Étant donné ces différents types de conditions et la déclaration de traits de la figure 7.6, on peut spécifier la dénotation de « ? » de nombreuses façons. Quelques exemples :

- `?[att3:v1]` dénote l'ensemble des nœuds qui ont le trait `att3:v1` ;
- `?[att3]` dénote l'ensemble des nœuds qui ont le trait `att3:v1` ou le trait `att3:v2` ou le trait `att3:v3` ;

- `?[att3:~]` dénote le complémentaire de l'ensemble précédent, c'est-à-dire l'ensemble des nœuds qui n'ont aucun trait ayant l'attribut `att3` ;
- `?[attG]` dénote l'ensemble des nœuds qui ont le trait `att1:+` ou le trait `att2:+` ou les deux.

Il est possible de formuler des conjonctions de conditions sur les nœuds (une virgule sépare les différentes conditions) :

- `?[attG,att3:v3]` dénote l'ensemble des nœuds qui ont le trait `att1:+`, ou le trait `att2:+`, ou ces deux traits, et qui ont également le trait `att3:v3`.

Ces différents exemples illustrent le fait que les traits associés aux nœuds de l'arbre syntaxique permettent de spécifier ce que nous appellerons des « catégories » de nœuds, une catégorie étant un ensemble de nœuds spécifié par les traits qui sont ou ne sont pas associés à ses éléments.

Nous avons vu à travers les exemples qu'une catégorie pouvait être plus ou moins spécifique, c'est-à-dire spécifier un ensemble de nœuds plus ou moins grand (par exemple, la catégorie `?[attG,att3:v3]` est incluse dans la catégorie `?[attG]`). La catégorisation des nœuds telle qu'elle existe pour l'analyse syntaxique du français sera présentée dans la section 7.3. Par la suite, nous ferons référence à une catégorie par la seule conjonction entre crochets des traits qui la caractérise (p. ex. en utilisant simplement `[attG]` pour `?[attG]`).

7.2.4 Modes d'assignation des traits

Pour terminer cette présentation du principe général des traits dans le système XIP, nous mentionnons ici brièvement les principaux modes d'assignation des traits aux nœuds de l'arbre syntaxique.

Le premier mode d'assignation de traits consiste en une traduction des étiquettes fournies par l'analyseur morphologique. Un exemple d'analyse morphologique est présenté figure 5.3 page 182. Considérons une lecture quelconque d'une unité lexicale fournie dans cette analyse, par exemple :

```
les      le       +InvGen+PL+Def+Det
```

Cette analyse donnera lieu dans le système XIP (sur la base d'une source d'information déclarée dans un fichier dédié) à la création d'un nœud lexical dominant l'unité *les*, nœud qui aura l'étiquette `DET` et auquel seront associés les traits suivants :

```
DET[det:+,def:+,pl:+,masc:+,fem:+]
```

La spécification et la sémantique de ces traits sera explicitée plus loin. Intuitivement, ces traits signifient que le nœud en question est un déterminant défini pluriel qui peut être masculin ou féminin (présence conjointe des traits `masc:+` et `fem:+`, traduction de l'étiquette morphologique `+InvGen`, « invariable en genre »).

Un autre mode d'assignation de trait(s) est fourni par les règles XIP elles-mêmes. Il est possible, dans une règle XIP, en même temps qu'on fait référence à un nœud, d'assigner un trait à ce nœud ou de supprimer un trait déjà assigné à ce nœud, par des formules de la forme `[attribut=valeur]` ou `[attribut=~]`, respectivement. Par exemple, dans la version actuelle de l'analyseur du français, lorsqu'une règle identifie un syntagme nominal (nœud `NP`) comme ayant la fonction sujet, celui-ci se voit assigner le trait `fonc:fsubj`.

D'autres mécanismes d'assignation de traits existent dans XIP, mécanismes que nous ne décrirons pas ici dans la mesure où ils n'ont pas d'incidence notable sur l'information que nous utiliserons. À cet égard, la présentation des deux modes d'assignation de traits que sont l'analyse morphologique et l'assignation de trait dans une règle XIP ne vise pas à détailler ces mécanismes, mais simplement à introduire l'idée que les traits n'ont pas pour seul rôle de spécifier les catégories des unités lexicales.

7.3 Catégories utilisées pour l'analyse du français

Le système de traits décrit dans la section précédente permet de spécifier des catégories, une catégorie étant un ensemble de nœuds spécifié par les traits qui doivent ou non être associés aux éléments de cet ensemble.

La présente section décrit un certain nombre de traits qui sont associés aux nœuds des arbres syntaxiques décrivant la structure des phrases du français. On se limite à la description des traits qui seront effectivement utilisés par notre système de résolution des expressions pronominales.

La très grande majorité des attributs utilisés dans l'implantation de l'analyseur du français n'ont en fait qu'une seule valeur possible : `+`. Sauf mention contraire explicite, dans ce qui suit, les attributs mentionnés doivent être interprétés comme étant déclarés par une formule de la forme `attribut:{+}`.

Nous serons amenés dans certains cas à spécifier l'extension d'une catégorie lexicale (une catégorie lexicale étant un ensemble d'unités lexicales, par opposition à une catégorie non lexicale, qui correspond à un nœud non lexical de l'arbre syntaxique). Dans ce cas, les formes données dans l'extension de la catégorie le sont en faisant abstraction des variations de casse (majuscules/minuscules) et des formes élidées. Par exemple, lorsqu'on dit que la catégorie des pronoms clitiques accusatifs non réfléchis contient la forme *le*, on veut dire qu'elle contient aussi, entre autres, les formes *LE* et *l'*.

Enfin, avant d'en venir à la description des différentes catégories pertinentes pour notre système, signalons qu'un nœud d'étiquette `ETIQ` a systématiquement un trait `etiq:+` [4]. Par exemple, un nœud `NP` a le trait `np:+`, un nœud `ST` a le trait

[4] Les différentes étiquettes de nœud sont présentées page 222 pour les nœuds lexicaux et page 225 pour les nœuds non lexicaux.

`st:+`, un nœud `DET` a le trait `det:+`, etc. Il y a donc une certaine redondance entre les étiquettes et les traits associés aux nœuds, mais celle-ci n'est pas sans intérêt dans la mesure où les conditions sur les traits donnent la possibilité de spécifier négativement des ensembles de nœuds. On peut par exemple faire référence à l'ensemble des nœuds qui ne sont ni des verbes ni des syntagmes verbaux par la formule `?[verb:~,fv:~,iv:~,gv:~]`.

7.3.1 Catégorisation des noms

Dans la mesure où nous chercherons à relier des expressions pronominales à des syntagmes nominaux (plus précisément aux noyaux de ces syntagmes), les noms constituent une catégorie importante pour notre système de résolution.

On décrit plus particulièrement dans la présente section la catégorisation des noms suivants deux axes : d'une part, nous décrivons comment sont catégorisés les noms propres, d'autre part, nous décrivons l'information à valeur sémantique qui est associée aux noms.

Traitement des noms propres

Grevisse [37, §451] distingue noms propres et noms communs de la manière suivante :

> Le nom commun est pourvu d'une signification, d'une définition, et il est utilisé en fonction de cette signification.
>
> Le nom propre n'a pas de signification véritable, de définition ; il se rattache à ce qu'il désigne par un lien qui n'est pas sémantique, mais par une convention particulière.

Grevisse distingue en outre, parmi les noms propres, d'une part les « véritables noms propres » (essentiellement les noms de lieux et de personne), d'autre part, des noms propres formés à partir de « mots ayant une signification » mais qui sont employés « pour désigner, en faisant abstraction de leur signification ». Nous appellerons ces noms propres des « noms propres compositionnels », voulant dire par là que ces noms propres sont composés de mots (noms, adjectifs, prépositions, etc.) communs de la langue et en suivant la syntaxe générale de la langue.

Ces deux catégories se retrouvent de manière presque équivalente chez K. Jonasson [47, section 2.2.2], qui distingue les noms propres « purs » (correspondant aux « véritables noms propres » de Grevisse) et les noms propres « à base descriptive ou mixte » (p. ex. *l'Académie française* et *la rue Monge*, respectivement). Les « noms propres à base descriptive » de Jonasson correspondent à nos « noms propres compositionnels ». Les noms propres « mixtes » de Jonasson, lorsqu'ils sont composés d'un nom commun suivi d'un nom propre véritable en apposition (p. ex. *la tour Eiffel*) ne sont pas analysé par nous comme des noms propres (voir ci-dessous la description de l'analyse de ce type de syntagme).

Les suites de mots suivantes sont des noms propres compositionnels :

- *L'éducation sentimentale* [5]
- *le Mouvement contre le racisme et pour l'amitié entre les peuples*
- *le Fonds monétaire international*
- *le Crédit agricole*

Le mot *Paris* ou la suite de mots *Jacques Chirac* sont de véritables noms propres [6]. Les noms propres étrangers qui seraient compositionnels dans la langue à la laquelle ils appartiennent (ex. *Banca di Roma, United Bank of Switzerland*) sont considérés par nous comme des noms propres véritables lorsqu'ils apparaissent dans des textes en français.

Noms propres véritables. Les noms propres véritables sont décrits dans l'arbre syntaxique par un nœud `NOUN` qui domine toute la séquence d'unités lexicales qui composent le nom propre. Ces nœuds ont le trait `proper:+`. Ils sont toujours noyau d'un syntagme nominal.

La phrase suivante contient trois noms propres véritables, dont la structure est représentée figure 7.7. Les nœuds `NOUN` dans chacun de ces trois fragments d'arbres ont le trait `proper:+`.

(4) « Nous nous sommes sentis relativement petits quand sont nés Citigroup ou l'United Bank of Switzerland », explique Maurice Lippens.

Lorsqu'un nom propre véritable est placé en apposition sans virgule à droite d'un nom commun (voir [37, §335]), comme, par exemple, dans les séquences suivantes :

- le prix Nobel
- l'affaire Greenpeace
- le président Chirac

de telles séquences sont analysées comme deux syntagmes nominaux distincts, liés entre eux par une relation de dépendance (voir ci-après section 7.4). Sont analysées de la même manière (c'est-à-dire comme deux syntagmes distincts) les séquences composées d'un titre (p. ex. *Monsieur, M., Mme*) suivi d'un nom propre véritable. Le nœud `NOUN` qui domine l'unité lexicale qui est un titre a le trait `tit:+`.

Le type de syntagme décrit ici correspond à ce que K. Jonasson appelle des « noms propres mixtes » [47, p. 36]. L'analyse que nous proposons est arbitraire ; elle est justifiée par le fait qu'il est difficile de déterminer une limite autre que celle qui passe par le choix de considérer tous ou aucun de ces syntagmes comme des

[5] Exemple de Grevisse.

[6] On voit avec cet exemple que « compositionnel » ne veut pas dire « composé de plusieurs mots », mais bien « composé de plusieurs mots communs ».

```
NP{NOUN{Citigroup}

NP{DET{l'} NOUN{United Bank of Switzerland}}

NP{NOUN{Maurice Lippens}}
```

FIG. 7.7 – Structure des noms propres véritables

```
NP{DET{le} NOUN{prix}} NP{NOUN{Nobel}}

NP{NOUN{M.}} NP{NOUN{Jacques Chirac}}
```

FIG. 7.8 – Noms propres véritables apposés

```
NP{DET{L'} NOUN{éducation}} AP{ADJ{sentimentale}}

NP{DET{le} NOUN{Fonds}} AP{ADJ{monétaire}} AP{ADJ{international}}
```

FIG. 7.9 – Structure des noms propres compositionnels

```
NP[pnhead:+]{DET[pnhead:+]{le} NOUN[pnhead:+]{Fonds}}
AP[pnpart:+]{ADJ[pnpart:+]{monétaire}}
AP[pnpart:+]{ADJ[pnpart:+]{international}}
```

FIG. 7.10 – Traits associés aux noms propres compositionnels

noms propres. Ainsi, face à des syntagmes tels que *l'action Paribas*, *le dossier GAN*, *le président Chirac*, *la méthode Jospin*, *la tour Eiffel*, nous ne sommes pas en mesure de spécifier un critère qui permette de déterminer lesquels sont des noms propres, lesquels n'en sont pas. Notre choix est donc de considérer que, dans de tels syntagmes, seul le mot commençant par une majuscule est un nom propre. Notons que ce choix n'a pas d'implication fondamentale pour notre système d'interprétation des pronoms : qu'un syntagme nominal tel que *la tour Eiffel* soit considéré dans son entier comme un nom propre ou comme un syntagme nominal noyau défini suivi d'un nom en apposition ne changerait rien dans le système qui sera décrit par la suite. Il s'agit ici de fixer la terminologie qui sera utilisée et le type de structure qu'elle décrit.

La figure 7.8 donne deux exemples de structures mettant en jeu un nom propre apposé. Les deux nœuds `NOUN` qui dominent respectivement *Nobel* et *Jacques Chirac* ont le trait `proper:+`. Le nœud `NOUN` qui domine *M.* a le trait `tit:+`.

Noms propres compositionnels. Alors que les noms propres véritables sont dominés immédiatement par un nœud `NOUN`, les noms propres compositionnels sont traités comme le sont les suites de syntagmes noyau en général. La figure 7.9 donne deux exemples de la structure des noms propres compositionnels.

Il est à noter qu'aucun nœud ne domine spécifiquement une suite de syntagmes noyau constituant un nom propre compositionnel. Nous avons cependant réimplanté dans le formalisme XIP les règles pour l'identification des noms propres que nous avions définies par le passé (voir [87]) de manière à identifier les noms propres compositionnels. La trace de cette identification n'apparaît pas sous la forme d'un nouveau nœud dans l'arbre syntaxique, mais sous la forme d'une assignation de trait aux expressions qui font partie d'un nom propre compositionnel [7]. Un nom propre compositionnel est vu comme une séquence de nœuds `N1 N2 ... Nn`. Par exemple, pour la suite *le Fonds monétaire international*, cette séquence est `NP AP AP` (voir figure 7.9). On appelle le nœud initial la « tête » du nom propre compositionnel. En l'état actuel de l'implantation, l'identification d'un nom propre compositionnel résulte en l'assignation du trait `pnhead:+` (« proper name head », c'est-à-dire « tête de nom propre ») au nœud tête du nom propre et à tous les nœuds dominés par ce nœud, et en l'assignation du trait `pnpart:+` (« proper name part », c'est-à-dire « partie de nom propre ») aux autres nœuds de la séquence qui constitue le nom propre compositionnel, ainsi qu'aux nœuds dominés par ces nœuds. Pour la suite *le Fonds monétaire international*, on aura donc l'association des traits illustrée figure 7.10.

Aucun des nœuds qui constituent un nom propre compositionnel ne reçoit le trait `proper:+`, sauf s'il domine une unité lexicale qui constitue un nom propre

[7] Créer un nouveau nœud dans l'arbre syntaxique aurait conduit à modifier l'arbre de telle manière qu'il n'était pas garanti que les règles qui, dans le processus d'analyse syntaxique du français, font référence à l'arbre par la suite s'appliquent correctement.

véritable. Par exemple, le nœud `NOUN` qui domine *France* dans le nom propre compositionnel *Banque de France* a à la fois le trait `proper:+` et le trait `pnpart:+`.

Information à valeur sémantique

De manière générale, la description du sens des unités lexicales est un des problèmes les plus difficiles de la linguistique, encore plus si l'on considère la désambiguïsation du sens de ces unités en contexte (problème dont nous verrons un exemple ci-après). Nous avons cependant expérimenté dans notre système de résolution des expressions pronominales l'usage d'une information à valeur sémantique pour les noms. Celle-ci cependant sera assez limitée.

L'information à valeur sémantique que nous utiliserons se limite à la caractérisation :

- d'une classe de noms à « valeur de numéral »,
- d'une classe de noms de « fractions »,
- d'une classe de noms de « mesures »,
- d'une classe de noms de « dates »,
- d'une classe de noms de « lieux »,
- d'une classe de noms de « personnes ».

La description de ces différentes classes est l'objet de la présente section. À l'exception des noms à valeur de numéral, des noms de fractions et des noms de dates, l'information sémantique que nous utilisons est extraite d'un dictionnaire électronique appelé « AlethDic ». Avant d'en venir à la description des différentes classes, nous décrivons brièvement cette ressource.

AlethDic. Le Centre de recherche européen de Xerox dispose d'un lexique morpho-syntaxico-sémantique du français appelé « AlethDic » [41], dans sa version 1.5.4. Ce dictionnaire a été produit par Gsi-Erli [8]. Il contient une couche d'information « sémantique » que nous avons été amenés à utiliser pour les noms. Nous nous limitons ici à une description partielle du dictionnaire AlethDic, c'est-à-dire de la seule information que nous avons utilisée.

Les différents sens des noms sont décrits dans AlethDic par l'association de chaque nom à une ou plusieurs classes sémantiques, chaque association d'un nom n_i à une classe c_i visant à rendre compte d'un sens de n_i. Les classes sémantiques sont elles-mêmes organisées hiérarchiquement selon une relation de subsomption.

La figure 7.11 donne la hiérarchie des classes auxquelles les noms sont associés dans AlethDic. La hiérarchie présentée n'est pas la hiérarchie complète ; les pointillés sous une classe donnée indiquent que des sous-classes existent pour la classe en question (p. ex. il existe trois sous-classes à l'intérieur de la classe TEMPS).

[8] Aujourd'hui Lexiquest.

En regard de chaque nom de classe, sont donnés quelques exemples de noms appartenant à la classe. On voit que les noms n'appartiennent pas nécessairement à une classe de spécificité maximale (c'est-à-dire à une classe terminale dans la hiérarchie). Un certain nombre de noms, par exemple, sont simplement associés à la classe ABSTRAIT, ce qui signifie qu'ils appartiennent à cette classe, mais à aucune de ses sous-classes [9].

L'ambiguïté d'un nom est représentée par le fait qu'il appartient à plusieurs classes, chaque appartenance d'un nom à une classe correspondant à un sens du nom. Le nom *banque*, par exemple, a trois sens dans AlethDic : il appartient à la classe DOMACT (la banque est un domaine d'activité, p. ex. *il travaille dans la banque*), à la classe HUMAIN (*banque* dans le sens de « société », p. ex. *le Crédit agricole est une banque*), ou à la classe LIEU (p. ex. *Pierre va à la banque*). Notons que de manière générale, pour une occurrence d'un nom n_i dans un texte, nous ne seront pas en mesure de savoir dans lequel de ses sens ce nom est employé (p. ex. étant donné une occurrence de *banque*, rien ne nous permettra de dire si ce nom est employé pour décrire le domaine d'activité, une société ou un lieu).

Noms à valeur de numéral et noms de fractions. Les noms à valeur de numéral et les noms de fractions constituent un sous-ensemble des noms que Grevisse décrit comme accompagnés d'un pseudo-complément [37, §422].

On appelle « noms à valeur de numéral » l'ensemble des noms tels que *dizaine, centaine, millier, million*, etc., dont Grevisse [37, §422c] dit qu'ils sont « particulièrement proche des déterminants numéraux » (p. ex. *trois*). Les noms à valeur de numéral ont le trait `numeral:+`.

On appelle « noms de fractions » un ensemble de noms, qui dans un syntagme nominal au singulier, sont susceptibles de décrire la totalité ou une partie d'un ensemble et se rencontrent fréquemment dans une structure `DET N1 des N2`, avec `N1` le nom de fraction (p. ex. *la plupart des discours*). La liste exhaustive de ces noms dans notre système est la suivante : *majorité, plupart, ensemble, totalité, quasi-totalité, partie, moitié, tiers, quart, cinquième, dixième, centième, millième, millionième*. Les noms de fractions ont le trait `fraction:+`.

Les noms à valeur de numéral et les noms de fractions nous intéressent pour deux raisons : d'une part, ils pourront être repris par une expression pronominale au pluriel, d'autre part, dans une structure `DET N1 DE (DET) N2`, avec `N1` un nom à valeur de numéral ou un nom de fraction, l'information sur la sémantique du syntagme sera portée par le second nom (`N2`). On aura, par exemple, pour le syntagme *une dizaine de jours*, l'analyse suivante :

```
NP{DET{une} NOUN{dizaine}} PP{PREP{de} NP{NOUN{jours}}}
```

Étant donné cette analyse, le nom *dizaine* sera considéré comme représentant du

[9] La définition de la classe ABSTRAIT est la suivante : « Ensemble des notions universellement reconnues, non classables à un niveau inférieur. » [41, p. 57].

```
NOTION
    |→ ABSTRAIT                    sentiment, bizarre, hygiène
    |     |→ ENTITÉ                garantie, repoussoir, paire, attribut
    |     |     |→ DIRECTION       cap, bâbord, aplomb
    |     |     |→ POINT_CARDINAL  est, ouest
    |     |     |→ LETTRE          l, m, capitale, hiéroglyphe
    |     |     |→ CHIFFRE         soixante, retenue, binôme, duel
    |     |     |→ SYSTPENSÉE      existentialisme, idéalisme, optimisme
    |     |     |→ MESURE          tonne, ratio, stère, teneur
    |     |     |→ TEMPS           période, germination, octobre
    |     |           |→ ...
    |     |→ MENTEFACT             alibi, indélicatesse, avertissement
    |     |→ QUANTITÉ              cuvée, bouchon, abîme
    |     |→ PHÉNOMÈNE             dérive, gangstérisme, céphalée
    |     |     |→ ...
    |     |→ ACTIVITÉ              vente, descente, plongeon, visite
    |     |     |→ OPÉRATION       repassage, regard, défense
    |     |     |→ DOMACT          banque, poésie, stomatologie
    |     |→ ATTRIBUT              maigreur, actualité, reste, bonté
    |     |     |→ ...
    |→ CONCRET                     cadavre, chose, immun, charogne
          |→ ANIMÉ                 déesse, végétal, vertébré
          |     |→ NON-HUMAIN      ours, bananier, branche
          |           |→ ...
          |     |→ HUMAIN          société, assureur, banque
          |           |→ ...
          |→ INANIMÉ               forme, bâtiment, objet, altération, matière
                |→ LIEU            tombeau, construction, aéroport
                |     |→ ...
                |→ OBJET           bricolage, androïde, hagiographie
                |     |→ ...
                |→ SUBSTANCE       hormone, moutarde, serpentine
                |→ ALTERATION      dommage, engelure, rupture
                |→ FORME           ruban, sarcome, mousse
```

FIG. 7.11 – Hiérarchie des classes sémantiques dans AlethDic

syntagme *une dizaine de jours*. L'information sur la sémantique de ce syntagme (c'est-à-dire le type d'objet qu'il désigne) sera présente au niveau du nœud qui domine le nom *jours* (le pseudo-complément selon Grevisse).

NOMS DE MESURES. La classe des noms de mesures est déterminée à partir des informations de AlethDic. Sont des noms de mesures tous les noms qui appartiennent à la classe MESURE de ce dictionnaire (voir figure 7.11), classe définie comme suit [41, p. 60] : « désigne toute unité conventionnelle de quantification (ex. *centimètre, ohm, livre*). » Outre ces noms, nous incluons également dans la classe des noms de mesures les noms catégorisés dans AlethDic comme des noms de monnaies (classe MONNAIE, sous-classe de la classe OBJET, non représentée sur la figure 7.11, non définie autrement que par la hiérarchie dans [41]). Les noms de mesures ont le trait `measure:+`.

NOMS DE DATES. La classe des noms de dates a été définie par les auteurs de l'analyseur du français. Elle contient les noms de jours (*lundi, mardi*, etc.), les noms de mois (*janvier, février*, etc.) et un ensemble de noms tels que *décennie, matin, trimestre, minute, weekend, surlendemain, début, fin*, noms susceptibles de décrire une date ou une période de temps. Les noms de dates ont le trait `time:+`.

NOMS DE LIEUX. La classe des noms de lieux est déterminée comme l'union de l'ensemble des noms qui appartiennent à la classe LIEU de AlethDic et de l'ensemble des noms propres de lieu du lexique de l'analyseur syntaxique (p. ex. *Paris, France, Moscou* appartiennent à la classe des noms de lieux). Les noms de lieux ont le trait `place:+`.

NOMS DE PERSONNES. La classe des noms de personnes est caractérisée comme l'ensemble des noms qui appartiennent à la classe HUMAIN d'AlethDic, auxquels s'ajoutent les titres tels que *Monsieur, Mme, Me*. Il faut ici comprendre le terme « personne » dans un sens très large. La classe des noms de personnes inclut aussi bien les noms qui peuvent décrire des personnes physiques, que ceux qui peuvent décrire des personnes morales ou groupes de personnes. Ainsi les noms *syndicat, groupe, famille, musée, université*, par exemple, appartiennent à la classe des noms de personnes. Les noms de personnes ont le trait `person:+`.

L'usage que nous ferons de l'information à valeur sémantique présentée dans cette section apparaîtra avec la description de notre système de résolution des expressions pronominales. Il importe cependant de rappeler ici que l'information utilisée le sera sans désambiguïsation du sens des noms, si bien que les différents traits encodant l'information sémantique doivent se lire comme des indications d'une possibilité que tel nom appartienne à telle classe et non comme une certitude qu'il est effectivement employé dans le sens qui détermine son appartenance à cette classe. Pour cette raison, l'usage de l'information à valeur sémantique

pour les noms aura une valeur relativement exploratoire dans notre système d'interprétation des pronoms.

7.3.2 Catégorisation des pronoms

On donne ici les grandes lignes de la catégorisation des pronoms, ceux-ci pouvant être plus finement catégorisés par les traits de nombre, de genre ou de personne qui seront présentés plus loin. La catégorisation présentée ici est essentiellement celle de Grevisse [37], si on excepte les pronoms numéraux qui ne sont pas catégorisés comme pronoms dans le système d'analyse syntaxique du français que nous utilisons.

On distingue les grandes classes de pronoms suivantes. Pour chaque classe, on donne entre crochets les traits qui suffisent à la caractériser.

- les pronoms clitiques : `[clit:+]`, p. ex. *il, me, lui, tu, vous, elles, le, se* ;
- les pronoms disjoints : `[ton:+]` (`ton` veut dire « toniques », autre appellation employée pour les pronoms disjoints), p. ex. *lui, elle, moi, soi, lui-même* ;
- les pronoms démonstratifs : `[pron:+,dem:+]`, p. ex. *celui, celle-ci, ceux-là, ce, ceci, ça* ;
- les pronoms relatifs : `[rel:+]`, p. ex. *qui, dont, auquel, où* ;
- les pronoms interrogatifs : `[pron:+,int:+]`, p. ex. *qui, auquel, où, quelle* ;
- les pronoms possessifs : `[pron:+,poss:+]`, p. ex. *le sien, la tienne* ;
- les pronoms « quantifieurs » : `[pron:+,quant:+]`, c'est-à-dire *plusieurs, aucun, chacun, certains, beaucoup, un, l'un* et les formes féminines correspondantes ;
- les pronoms indéfinis : `[pron:+,indef:+]`, p. ex. *quiconque, quelqu'un, personne, d'aucun, autrui, tous, nul.*

Quelques remarques sur cette première catégorisation. Les traits `clit:+`, `ton:+` et `rel:+` ne sont associés qu'à des nœuds qui ont également le trait `pron:+`, si bien que ces traits suffisent à caractériser la classe en question comme étant une sous-classe des pronoms [10]. Les deux premières classes sont les plus importantes pour notre système puisqu'elles contiennent les formes dont nous devons spécifier l'interprétation. Nous spécifierons ci-après des sous-classes à l'intérieur de ces deux classes, ainsi qu'à l'intérieur de la classe des pronoms démonstratifs. La spécification de sous-classes pour les autres catégories définies ci-dessus ne se justifie pas pour notre système de résolution des expressions pronominales, compte tenu des expressions que nous avons choisi de traiter.

L'union de la classe des pronoms quantifieurs et de celle des pronoms indéfinis constitue *grosso modo* la classe des pronoms indéfinis de Grevisse ([37, §705]). Le

[10] Le trait `int:+` est également assigné à quelques formes qui ne sont pas des pronoms (p. ex. *quand, comment*).

terme « pronom indéfini » n'a donc pas ici la même extension que chez cet auteur. Notre distinction est basée sur un critère syntaxique : un pronom quantifieur est un pronom qu'on rencontrera fréquemment dans des structures de la forme :

```
PRONOM de DET NOM-PLURIEL
```

Exemples : *aucun des candidats, chacun des États, certains des invités*. Les pronoms indéfinis dans notre système de catégories n'ont pas cette propriété.

Notons pour finir que les pronoms numéraux ne sont pas catégorisés comme des pronoms par les auteurs de l'analyseur du français. Les formes telles que *deux, trois*, etc. sont simplement catégorisées comme des unités de catégorie `[num:+]`. L'emploi pronominal de ces formes est repérable par le fait que le numéral constitue le noyau d'un syntagme nominal comme dans l'arbre suivant :

```
ST{SC{NP{NUM{Deux}} FV{VERB{ont} VERB{démissionné}}} SENT{.}}
```

Catégorisation des pronoms clitiques

La figure 7.12 donne la catégorisation des pronoms clitiques de troisième personne [11], abstraction faite des traits rendant compte du genre et du nombre des pronoms. La figure montre un arbre dont la racine est `clit:+`. En suivant le chemin de la racine à un nœud terminal de l'arbre, on obtient l'ensemble des conditions sur les traits qui spécifient la classe indiquée à droite en *italiques*. Par exemple, la catégorie `[clit:+,refl:~,dat:+]` est constituée des deux formes *lui* et *leur*. On remarquera que la forme *se* est ambiguë en ce qui concerne les traits `acc:+` et `dat:+`.

Catégorisation des pronoms disjoints

La figure 7.13 présente la catégorisation des pronoms disjoints de troisième personne sur le même modèle que la figure 7.12. Cette catégorisation est limitée à une simple partition de la classe en deux sous-ensembles : les pronoms disjoints réfléchis et les pronoms disjoints non réfléchis.

Catégorisation des pronoms démonstratifs

Une catégorisation des pronoms démonstratifs nous intéresse dans la mesure où certains d'entre eux ne pourront *a priori* pas être source d'une reprise. Il s'agit des pronoms *ce, ceci, cela* et *ça*. Ces pronoms, à l'intérieur de la classe des pronoms démonstratifs, se distinguent des autres par le fait qu'ils sont de genre

[11] Les pronoms de première et deuxième personnes présentent peu d'intérêt ici, puisqu'ils sont exclus de notre objectif.

```
clit:+
|→ refl:~
|    |→ nom:+
|    |    |→ indef:~    {il, elle, ils, elles}
|    |    |→ indef:+    {l'on, on}
|    |→ acc:+           {le, la, les}
|    |→ dat:+           {lui, leur}
|    |→ gen:+           {en}
|    |→ loc:+           {y}
|→ refl:+
     |→ acc:+           {se}
     |→ dat:+           {se}
```

FIG. 7.12 – Catégorisation des pronoms clitiques de 3e personne

```
ton:+
|→ refl:~ {lui, elle, eux, elles, lui aussi, elle aussi, eux aussi, elles aussi}
|→ refl:+ {lui-même, elle-même, eux-mêmes, elles-mêmes, soi-même, soi}
```

FIG. 7.13 – Catégorisation des pronoms disjoints de 3e personne

```
ton:+
|→ masc:+ {ceux, ceux-là, ceux-ci, celui, celui-là, celui-ci}
|→ fem:+ {celles, celle, celle-là, celle-ci, celles-là, celles-ci}
|→ neutre:+ {ça, ce, ceci, cela, tout ce}
```

FIG. 7.14 – Catégorisation des pronoms démonstratifs

```
det:+
|→ def:+ {le, la, les, ledit, du ...}
|→ dem:+ {ce, cet, cette, ces, ...}
|→ poss:+ {son, sa, ses, leur, leurs, ...}
|→ indef:+ {un, une, de, des, tel, bon nombre de, ...}
|→ quant:+ {maints, tout, nul, chaque, aucun, ...}
|→ int:+ {quel, quelle, quels, quelles}
```

FIG. 7.15 – Catégorisation des déterminants

neutre. Avec les traits `masc:+`, `fem:+` et `neutre:+` pour rendre compte du genre des unités lexicales, on a les classes présentées figure 7.14.

7.3.3 Catégorisation des déterminants

Les déterminants sont répartis en six classes, présentées figure 7.15 : déterminants définis (`def:+`), démonstratifs (`dem:+`), possessifs (`poss:+`), indéfinis (`indef:+`), quantifieurs (`quant:+`) et interrogatifs (`int:+`). Outre les formes indiquées, les déterminants définis, démonstratifs et possessifs contiennent des formes composées avec *tout* (p. ex. *tout le, tous ces, toute sa*). La classe des déterminants possessifs contient également les formes de première et deuxième personnes correspondant aux formes de troisième personne indiquées (p. ex. *mon, toute ta*). Pour les déterminants définis, nous n'avons pas indiqué toues les variations de la forme *ledit* (p. ex. *lesdits, lesdites*). Enfin la classe des déterminants quantifieurs contient des déterminants complexes composés d'un adverbe ou d'un pronom quantifieur suivi de la préposition *de* dans le contexte d'un nom pluriel à droite sans déterminant. Par exemple, dans *beaucoup de gens*, la suite *beaucoup de* est considérée comme un déterminant quantifieur ; par contre, dans *beaucoup de ces personnes*, *beaucoup* est un pronom.

Les prépositions (nœuds `PREP`) intégrant le déterminant défini contracté ont les traits `prep:+` et `def:+`, mais n'ont pas le trait `det:+`. Il s'agit des formes *au, aux, du, des*, d'un certain de nombre de prépositions complexes se terminant par une de ces formes (p. ex. *en dépit des, jusqu'au*) et des formes *dudit, desdites*, etc.

Deux attributs généraux référencent respectivement les déterminants définis, démonstratifs et possessifs, qui sont dits « déterminés » (attribut général `deter`), et les déterminants quantifieurs et indéfinis, qui sont dits « indéterminés » (attribut général `indeter`). On a la déclaration de traits présentée figure 7.16.

```
deter:[
  def:{+},
  dem:{+},
  poss:{+}],
indeter:[
  quant:{+},
  indef:{+}],
int:{+}
```

FIG. 7.16 – Déclaration des traits associés aux déterminants.

Les numéraux (p. ex. *deux, trois*, étiquette `NUM`, catégorie `[num:+]`) ne sont jamais considérés comme de catégorie `[det:+]`. Ils pourrons néanmoins occuper la fonction d'un déterminant.

7.3.4 Nombre, genre et personne

Traits rendant compte du nombre.

Nous distinguons le nombre « grammatical », qui est le nombre tel qu'on l'entend habituellement, et le nombre « sémantique », qui est lié à la dénotation des expressions. L'intérêt et la sémantique de cette distinction apparaîtra plus clairement ci-après avec la description du nombre grammatical et du nombre sémantique des déterminants possessifs.

NOMBRE GRAMMATICAL. L'information sur le nombre grammatical des expressions est codée par les deux traits suivants :

```
sg:+
pl:+
```

Une unité lexicale (nom, verbe, adjectif, déterminant, pronom, etc.) de nombre singulier a le trait `sg:+`. Une unité lexicale de nombre pluriel a le trait `pl:+`.

Étant donné nos objectifs, les traits de nombre grammatical nous intéresseront essentiellement pour les déterminants possessifs, les pronoms et les expressions qui sont le noyau d'un syntagme nominal.

NOMBRE SÉMANTIQUE. En règle générale, une expression pronominale s'accorde avec sa source suivant leur nombre grammatical. Par exemple, dans la phrase

(5) Le président a dit qu'il ne démissionnerait pas.

le pronom *il* et *le président* sont deux expressions de nombre grammatical singulier.

Cette observation ne vaut pas pour tous les cas de reprise pronominale, en particulier les reprises par un déterminant possessif. Dans la phrase

(6) Le président a présenté ses objectifs.

le déterminant possessif *ses* est de nombre grammatical pluriel et sa source (*Le président*) est de nombre grammatical singulier.

En règle générale, nous pouvons dire pour les déterminants possessifs que *son, sa* et *ses* renvoient à une expression de nombre grammatical singulier et que *leur* et *leurs* renvoient à une expression de nombre grammatical pluriel. Nous dirons des trois premières expressions qu'elles ont un nombre sémantique singulier et des deux dernières qu'elles ont un nombre sémantique pluriel, information dont on rend compte par deux traits : `semsg:+` et `sempl:+`. Pour les déterminants possessifs de troisième personne, on aura donc les catégories suivantes :

`[det:+,poss:+,sg:+,semsg:+]` = *{son, sa}*
`[det:+,poss:+,pl:+,semsg:+]` = *{ses}*
`[det:+,poss:+,sg:+,sempl:+]` = *{leur}*
`[det:+,poss:+,pl:+,sempl:+]` = *{leurs}*

Par « expression de nombre sémantique singulier (resp. pluriel) », nous voulons donc dire « expression dénotant *a priori* un être singulier (resp. un ensemble d'êtres). En dehors des déterminants possessifs, il y a en général une correspondance directe entre le nombre grammatical et le nombre sémantique : un syntagme nominal grammaticalement singulier dénote en général un être singulier, un syntagme nominal pluriel dénote en général un ensemble d'êtres. Compte tenu de cette régularité, nous utiliserons les traits `semsg:+` et `sempl:+` principalement pour rendre compte des cas qui échappent à la règle générale.

Dans le système de résolution des expressions pronominales tel que nous l'avons implanté, reçoivent un trait de nombre sémantique, outre les déterminants possessifs déjà évoqués, les formes suivantes :

- le trait `sempl:+` est associé aux noms de fractions et aux noms à valeur de numéral (voir p. 238) ;
- le trait `sempl:+` est associé aux expressions qui sont le noyau d'un syntagme nominal déterminé par *aucun* [12].

Pour permettre une application unifiée des règles qui seront présentées au chapitre 10, les traits `semsg:+` et `sempl:+` sont également associés aux expressions de catégorie `[pron:+]` qui ont le trait `sg:+` et `pl:+`, respectivement.

Traits rendant compte du genre

Nous avons déjà vu plus haut les traits rendant compte du genre (« Catégorisation des pronoms démonstratifs », p. 243). Ils sont au nombre de trois :

`masc:+` (masculin)
`fem:+` (féminin)
`neutre:+` (neutre)

Le trait `neutre:+` n'est associé qu'aux pronoms démonstratifs déjà évoqués. Une forme qui est ambiguë quant au genre a à la fois le trait `masc:+` et le trait `fem:+` C'est le cas, par exemple, des pronoms clitiques *lui* et *leur*.

Traits rendant compte de la personne

Les traits rendant compte de la personne sont :

`p1:+` (première personne)
`p2:+` (deuxième personne)
`p3:+` (troisième personne)

[12] Dans la représentation des relations syntaxiques sous forme de dépendances, les syntagmes sont représentés par leur noyau (voir section 7.4).

Ils nous intéresserons dans la mesure où ils nous permettrons d'exclure les expressions pronominales de première et deuxième personne de l'ensemble des reprises et de l'ensemble des sources possibles.

7.3.5 Prépositions

Quelques prépositions auront un intérêt plus particulier dans notre système de résolution des expressions pronominales. De manière générale, les différentes prépositions sont caractérisées par un trait ayant l'attribut `form`, qui peut prendre de nombreuses valeurs, chacune spécifiant une préposition ou une classe de prépositions particulière. Nous donnons ici les traits associés aux prépositions auxquelles il sera fait référence dans le système de résolution.

`form:fde` = *{de, du, des}*
`form:fdentre` = *{d'entre}* (p. ex. dans *deux d'entre eux*)
`form:fen` = *{en}*
`form:fa` = *{à, au, aux}*
`form:fpar` = *{par}*
`form:fquanta` = *{quant à/au/aux, pour ce qui est de/du/des}*

Outre ces traits qui spécifient des classes très réduites, nous ferons également usage du trait `locprep:+` qui caractérise la classe des prépositions ayant une valeur locative, à savoir, dans l'implantation actuelle, l'ensemble suivant :

{autour de, dans, devant, derrière, en, en dehors de, face à, sous, sur, vers}

7.3.6 Formes verbales

L'information que nous utiliserons, en ce qui concerne les formes verbales, est relativement limitée. Sont distingués les verbes à valeur de copule, une classe de verbes que nous appelons « verbes de discours » et les formes à l'infinitif.

Copules

Dans l'ensemble des verbes, nous distinguons deux sous-classes sémantiques, les verbes à valeurs de copules et les verbes de discours.

Les verbes à valeur de copule sont les suivants :

{être, apparaître, demeurer, devenir, paraître, redevenir, rester, sembler}

Les différentes formes de ces verbes apparaissant dans le texte analysé ont le trait `copule:+`. À l'intérieur de cet ensemble, les formes du verbe *être* sont en outre distinguées par le trait `form:fetre`.

Verbes de discours

On appelle « verbes de discours » un ensemble de 132 verbes caractérisés comme pouvant être noyau d'une proposition incise avec sujet à droite, la proposition exprimant le fait que l'être dénoté par son sujet est l'auteur du discours ou la source de l'information rapporté(e) dans la phrase où apparaît l'incise.

Dans le texte suivant :

(7) Le groupe Paribas va céder la participation de 25 % qu'il détient dans la banque d'affaires russe United Financial Group (UFG) au management de cette dernière. Cette décision, précise-t-il, lui permettra de conduire le développement de ses activités en Russie dans le cadre de son organisation mondiale par métier.

la proposition incise *précise-t-il* exprime le fait que l'information selon laquelle la décision en question permettra à Paribas de conduire, etc. a pour source l'être dénoté par *il*, à savoir le groupe Paribas.

Cet ensemble de 132 verbes a été extrait à partir d'un corpus d'articles du *Monde* (un peu plus de 6 millions de mots), sur la base d'expressions régulières spécifiant des structures typiques des incises (p. ex. une virgule suivie d'un syntagme verbal noyau, suivi d'un pronom sujet). Les verbes extraits ont été vérifiés manuellement.

Les verbes de discours ont le trait `dicendi:+`. Voici un échantillon des verbes catégorisés comme verbes de discours dans notre système :

{conseiller, dire, ordonner, proposer , recommander, suggérer, expliquer, ajouter, déclarer, affirmer, estimer, souligner, écrire, poursuivre, ...}

Formes à l'infinitif

De manière générale, aux formes verbales sont associés des traits qui rendent compte du temps (présent, futur, etc.) et du mode (indicatif, subjonctif, etc.). Dans notre système de résolution des expressions pronominales, nous ne ferons référence qu'à un seul de ces traits : le trait `inf:+`, qui est associé aux formes verbales à l'infinitif.

7.3.7 Propositions

On rappelle que l'arbre syntaxique produit par le système d'analyse du français ne contient pas de nœuds qui domineraient des propositions complètes, mais plutôt des nœuds qui dominent des propositions finies noyau. Ces nœuds sont étiquetés `SC`.

En sortie de l'analyseur syntaxique, les propositions incises avec sujet à droite et dont le verbe noyau est un verbe de discours (voir ci-dessus section 7.3.6) sont distinguées des autres propositions par le fait qu'elles ont le trait `disco:+`.

```
ST{SC{NP{DET{Cette} NOUN{décision}}
      PUNCT{,}
      SC{FV{VERB{précise}}}
      NP{PRON{-t-il}}
      PUNCT{,}
      FV{PRON{lui} VERB{permettra}}}
   IV{PREP{de} VERB{conduire}}
   NP{DET{le} NOUN{développement}}
   PP{PREP{de} NP{DET{ses} NOUN{activités}}}
   PP{PREP{en} NP{NOUN{Russie}}}
   SENT{.}}
```

FIG. 7.17 – Exemple d'arbre syntaxique (3)

La figure 7.17 présente l'arbre construit pour la phrase *Cette décision, précise-t-il, lui permettra de conduire le développement de ses activités.* Cette arbre contient deux nœuds `SC`, un qui domine la suite d'unités lexicales *Cette décision, précise-t-il, lui permettra*, un autre qui domine seulement *précise*. Ce dernier a le trait `disco:+`.

7.3.8 Phrases

On distingue parmi les nœuds `ST` (les « phrases », voir section 7.1.3) ceux qui ne dominent immédiatement aucun nœud d'étiquette `SC` (proposition noyau), `IV` (syntagme verbal infinitif) et `GV` (syntagme verbal avec forme au participe présent). Ces nœuds `ST` ont le trait `noverb:+`. Nous dirons des séquences d'unités lexicales dominées par de tels nœuds qu'elles constituent des « phrases sans verbe ».

Dans l'arbre syntaxique de la figure 7.3 page 227, arbre décrivant la suite *Autre candidat à devoir encore faire ses preuves : Eureko.*, le nœud `ST` qui domine la suite :

Eureko.

a le trait `noverb:+`, mais pas le nœud `ST` qui domine la suite :

Autre candidat à devoir encore faire ses preuves :

car ce nœud domine immédiatement un nœud `IV`.

Les phrases sans verbe sont le plus souvent dans notre corpus des segments introduits à gauche par « : », ou les titres ou intertitres que l'on rencontre dans les articles de journaux [13].

[13] Le texte en entrée du système ne contient pas d'information explicite (p. ex. de type HTML) sur sa structure en titre, paragraphe, etc.

7.3.9 Insertions

Il est possible d'isoler à l'intérieur d'une phrase des segments qui peuvent être vus comme « insérés » dans le texte pour y apporter des précisions sur ce qui est dit par ailleurs dans la phrase. Nous appelons ces segments de textes des « insertions ». De manière générale et approximative, nous caractériserions une insertion comme un segment de texte :

- délimité à gauche par le début de la phrase ou un symbole de ponctuation,
- délimité à droite par la fin de la phrase ou un symbole de ponctuation,
- qui peut être supprimé sans nuire à la correction grammaticale de la phrase,
- et qui apporte une précision par rapport au discours principal constitué par la phrase sans l'insertion.

L'exemple le plus évident d'insertion se rencontre avec l'usage des parenthèses. Dans le texte suivant,

(8) L'intersyndicale du Crédit Foncier a obtenu gain de cause. Elle sera reçue par Dominique Strauss-Kahn le 20 mai prochain et le 26 mai par Jean-Claude Gayssot, le ministre des Transports et du Logement. Les négociations se sont déroulées tandis que près de 140 salariés du Foncier occupaient les mairies de Sarcelles (Dominique Strauss-Kahn est maire adjoint) et de Drancy (Jean-Claude Gayssot est conseiller municipal).

les deux segments entre parenthèses sont des insertions : ils peuvent être supprimés et apportent une précision expliquant le fait que les mairies de Sarcelles et Drancy (précisément elles et pas d'autres mairies) soient occupées par les salariés du Foncier.

Dans cet exemple, le segment *le ministre des Transports et du Logement* est également une insertion. Il est délimité à gauche par une virgule, à droite par la fin de la phrase, peut être supprimé et exprime une précision sur la fonction de Jean-Claude Gayssot.

L'information véhiculée par les insertions a dans le texte un caractère secondaire. Pour cette raison, en règle générale, une expression pronominale ne renverra pas à une expression qui se trouve dans une insertion, sauf éventuellement, si cette expression pronominale se trouve elle-même dans l'insertion [14]. Les insertions présentent donc un intérêt certain pour notre système de résolution des expressions pronominales.

On distingue deux types d'insertions : les insertions entre parenthèses, crochets (« [ ] »), accolades (« { } ») ou tirets, d'une part, et des insertions mettant

[14] La notion d'insertion que nous utilisons n'est pas sans rappeler, toute proportion d'échelle gardée, la notion d'unités de discours satellites telle que définie dans la théorie des structures rhétoriques et l'observation que nous formulons ici au sujet des insertions n'est pas sans rapport avec les hypothèses de la théorie des veines (voir la section 6.4.2 du chapitre 6, qui est consacrée à cette dernière).

en jeu une virgule à gauche et/ou à droite, d'autre part. Ces dernières sont restreintes à quelques contextes particuliers, décrits ci-après. L'information selon laquelle une expression fait partie d'une insertion sera codée dans les deux cas par l'association d'un trait : `inser:+` dans le cas des insertions du premier type, `embed:+` dans le cas des insertions du second type.

Insertions entre parenthèses

Dans l'arbre syntaxique construit pour les textes en français, les segments entre parenthèses, crochets, accolades ou tirets sont dominés par des nœuds d'étiquette `INS`.

L'arbre syntaxique présenté figure 7.18, analyse de la phrase suivante,

(9) Dominique Strauss-Kahn a donné son accord pour une prorogation des mandats des instances dirigeantes du groupe — c'est-à-dire du conseil de surveillance du Cencep — jusqu'à la promulgation de la loi de réforme.

contient un nœud `INS` qui domine la séquence d'unités lexicales :

— *c'est-à-dire du conseil de surveillance du Cencep* —

Tout nœud dominé, immédiatement ou non, par un nœud `INS` reçoit le trait `inser:+`. C'est donc le cas, par exemple, pour les nœuds `NOUN` dominant les unités lexicales *conseil, surveillance* et *Cencep* dans l'arbre de la figure 7.18.

Insertions avec virgule(s)

Le deuxième type d'insertion que nous considérons est marqué par la présence d'une virgule à gauche et/ou à droite de l'insertion. Ces insertions sont identifiées dans les contextes décrits ci-dessous. Les nœuds qui dominent une séquence d'unités lexicales entièrement incluse dans une insertion ont le trait `embed:+`.

Il importe de noter que la définition des insertions que nous avons proposée ci-dessus au début de la section 7.3.9 doit être lue comme une définition approximative qui vise à donner une idée de ce qu'on entend par ce terme. En pratique, pour les insertions avec virgules, nous n'avons pas menée une étude systématique des insertions dans le sens de cette définition, mais plutôt défini des contextes syntaxiques nous permettant d'identifier *certaines* insertions. Ces contextes sont partie intégrante de la définition des insertions entre virgules et c'est au regard de cette définition plus spécifique que le système devra être évalué.

Cela étant, les différents contextes définissant les insertions avec virgules sont les suivants (les chiffres entre parenthèses renvoient aux exemples donnés plus loin) :

- insertion entre le sujet `X` d'un verbe fléchi `Y` et le verbe `Y`. L'insertion est délimitée à gauche par une virgule et à droite par une virgule ou une insertion entre parenthèses (10-12) ;

```
ST{SC{NP{NOUN{Dominique Strauss-Kahn}}
      FV{VERB{a} VERB{donné}}}
   NP{DET{son} NOUN{accord}}
   PP{PREP{pour} NP{DET{une} NOUN{prorogation}}}
   PP{PREP{des} NP{NOUN{mandats}}}
   PP{PREP{des} NP{NOUN{instances}}}
   AP{ADJ{dirigeantes}}
   PP{PREP{du} NP{NOUN{groupe}}}
   INS{
      PUNCT{-}
      COORD{c'est-à-dire}
      PP{PREP{du} NP{NOUN{conseil}}}
      PP{PREP{de} NP{NOUN{surveillance}}}
      PP{PREP{du} NP{NOUN{Cencep}}}
      PUNCT{-}}
   PP{PREP{jusqu'à} NP{DET{la} NOUN{promulgation}}}
   PP{PREP{de} NP{DET{la} NOUN{loi}}}
   PP{PREP{de} NP{NOUN{réforme}}}
   SENT{.}}
```

FIG. 7.18 – Exemple d'arbre syntaxique (4).

- insertion entre un verbe fléchi, infinitif ou participe présent et son complément d'objet direct (13-14). Cette règle est restreinte aux compléments d'objets nominaux [15] ;
- appositions à droite (15) ;
- appositions à gauche, à la condition qu'elles portent sur le sujet (16) [16].

Dans les exemples suivants, sont indiqués en *italiques* les segments de texte considérés comme des insertions avec virgule(s).

(10) Les discussions entre la chancellerie et les greffiers des tribunaux de commerce sur l'abaissement du tarif d'accès à leur serveur Minitel (La Tribune du 23 février), *qui avaient démarré au début de l'année sur les chapeaux de roue,* piétinent.

(11) Robert Panhard, *cinquante-deux ans,* a été élu président de la Chambre des notaires de Paris.

(12) L'investissement, *qui se chiffre à 28,6 milliards de lires (un peu moins de 100 millions de francs)* est revenu à racheter 10 % à la Banca Agricola Mantovana.

[15] Par opposition aux compléments propositionnels ou verbaux.

[16] L'exemple (17) illustre le cas où une apposition gauche n'est pas une insertion car elle ne porte pas sur le sujet.

(13) Présentant jeudi dernier le rapport annuel de la cour suprême retraçant, *à l'intention du garde des Sceaux,* son activité de l'année 1997, le procureur général de la Cour de cassation a souligné cette tendance de fond qui tend désormais à « mettre fin à l'exception française ».

(14) Au rayon des statistiques, le rapport établit, *après une correction d'inventaire,* à 38 452 le nombre de dossiers en attente au 31 décembre 1997.

(15) Selon Jean Coroller, *associé d'Ernst & Young Audit et directeur du département de contrôle interne,* « une implication personnelle et active des administrateurs et des dirigeants est de nature à diminuer les risques ».

(16) Le cabinet d'avocat Mazars & Associés s'est rapproché du cabinet lyonnais Michaud. *Fondé en 1981 par Pierre-Henry Michaud,* cette structure compte 3 avocats. *Spécialisé en droit des sociétés,* il vient compléter l'activité de droit social de Mazars & Associés à Lyon.

Dans l'exemple suivant, le segment *Seul regret pour Swiss Life* peut être vu comme une apposition à gauche, mais il n'est pas une insertion selon notre définition, car il ne porte pas sur le sujet mais plutôt sur l'ensemble de la proposition qui le suit (le regret, c'est que le CCF n'ait pas été retenu). En revanche, les deux segments en italiques sont des insertions.

(17) Seul regret pour Swiss Life, le CCF, *dont il est actionnaire,* n'a pas été retenu pour le CIC, *partenaire de bancassurance du GAN*.

7.3.10 Autres traits utilisés

Pour terminer cette présentation des différents traits qui seront utilisés dans notre système de résolution, nous regroupons dans cette dernière section deux traits particuliers, assignés respectivement aux nœuds lexicaux qui dominent une virgule et aux syntagmes nominaux sujet.

Le trait `form:fcm` est associé aux nœuds d'étiquette `PUNCT` qui dominent une virgule.

Le trait `fonc:fsubj` est associé aux syntagmes nominaux noyau (nœuds `NP`) sujet.

7.4 Dépendances

Outre l'arbre syntaxique partiel que nous venons de décrire, l'analyseur XIP du français propose une description de la structure syntaxique des phrases en termes de « dépendances ». Plus que l'arbre syntaxique, dont nous avons dit qu'il n'était qu'un moyen facilitant l'identification des dépendances, ce sont les dépendances qui visent à décrire complètement les phrases analysées. L'objet

de la présente section est de décrire les différentes dépendances extraites par l'analyseur syntaxique du français.

En complément des exemples qui seront présentés ici, l'annexe B présente l'ensemble des dépendances syntaxiques extraites par l'analyseur syntaxique du français pour un court extrait d'un article de *La Tribune*.

7.4.1 Des relations entre les nœuds de l'arbre syntaxique

De manière informelle, nous verrons ici les dépendances produites par l'analyseur syntaxique comme des relations sur des nœuds de l'arbre syntaxique. Ces relations sont exprimées par des prédicats de la forme

`PREDICAT(arg_1,arg_2,...,arg_n)`

où `arg_1, arg_2, ..., arg_n` font référence à des nœuds de l'arbre syntaxique et `PREDICAT` est le nom de la relation qui existe entre les nœuds. Dans le formalisme XIP, les relations peuvent avoir un nombre quelconque d'arguments. En particulier, il est tout à fait possible de formuler une relation unaire, c'est-à-dire une relation ayant un seul argument.

7.4.2 Inventaire des relations syntaxiques

On fait ici l'inventaire des relations syntaxiques qui seront utilisées par notre système d'interprétation des expressions pronominales. Pour indiquer l'arité des différentes relations et faire référence aux différents arguments, on utilise les variables `X`, `Y` et `Z`. Pour chaque relation, différents exemples sont donnés. Les chiffres entre parenthèses dans le corps du texte font référence à ces exemples. Par convention, on représente les nœuds arguments d'une relation par la suite d'unités lexicales qu'ils dominent, c'est-à-dire les mots tels qu'ils apparaissent dans le texte analysé [17].

Le verbe et son (ou ses) sujet(s)

• `subj(X,Y)`

`Y` est sujet de `X`, `X` étant le noyau d'un syntagme verbal noyau. `Y` est soit le noyau d'un syntagme nominal noyau (18), soit, beaucoup plus rarement, le noyau d'un syntagme verbal (19).

(18) Jacques dort.
`subj(dort,Jacques)`

[17] Il va sans dire que chaque nœud est dûment caractérisé de manière univoque dans le système et, si l'usage des unités lexicales pour représenter les nœuds pourra donner lieu pour le lecteur à une ambiguïté (dans le cas ou deux nœuds distincts dominent des unités lexicales de même forme), celle-ci ne sera qu'apparente et due à la notation.

(19) Dormir lui fait du bien.
`subj(fait,Dormir)`

Dans le cas où le sujet est une coordination de syntagmes, on a une relation `subj` pour chaque noyau de chacun des syntagmes coordonnés (20). L'arbre syntaxique ne contient pas de nœud qui domine l'ensemble des syntagmes coordonnés.

(20) Pierre, Jacques et Jean sont venus.
`subj(venus,Pierre)`
`subj(venus,Jacques)`
`subj(venus,Jean)`

Dans le cas où on a plusieurs verbes coordonnés, avec le sujet exprimé une fois seulement avant le premier verbe, on a autant de relations `subj` qu'il y a de verbes (21).

(21) Jacques mange et boit.
`subj(mange,Jacques)`
`subj(boit,Jacques)`

L'analyseur syntaxique rend également compte au moyen de la relation `subj` du contrôle des verbes à l'infinitif (22). Les syntagmes verbaux infinitifs ou participes ont un argument sujet implicite, qu'on dit « contrôlé » par un constituant de la phrase. Ce dernier est alors dit « contrôleur » du sujet implicite du verbe dont le sujet est implicite [18]. L'analyseur syntaxique identifie une relation `subj` entre un verbe dont le sujet est implicite et le noyau du syntagme qui contrôle le sujet implicite de ce verbe. Dans l'exemple (22), les verbes *renforcer*, *obliger* et *déposer* ont tous les trois un sujet implicite contrôlé pour les deux premiers par le syntagme *Ce décret* et pour le troisième par le pronom *les*. Trois relations `subj` rendent compte de ces trois phénomènes de contrôle.

(22) Ce décret vise à renforcer les contrôles sur cette profession et en particulier à les obliger à déposer leurs fonds auprès de la Caisse des dépôts.
`subj(vise,décret)`
`subj(déposer,les)`
`subj(renforcer,décret)`
`subj(obliger,décret)`

L'analyseur syntaxique identifie également une relation `subj` entre un verbe au participe présent dont le sujet est implicite et le syntagme qui contrôle ce sujet (23).

(23) Il arrive en boitant.
`subj(arrive,Il)`
`subj(boitant,Il)`

[18] Sur ce sujet, voir, par exemple, Karine Baschung. *Grammaires d'unification à traits et contrôle des infinitives en français.* Adosa, Clermont-Fd, 1991. p. 15-17.

Dans le cas où le sujet d'un verbe `X` est le pronom relatif *qui*, l'analyseur produit deux relations `subj` avec `X` en premier argument : une pour le pronom relatif et une pour l'antécédent du pronom relatif. Cet antécédent peut être par ailleurs sujet d'un autre verbe (24) [19].

(24) Les discussions, qui avaient démarré sur les chapeaux de roue, piétinent.
```
subj(démarré,qui)
subj(piétinent,discussions)
subj(démarré,discussions)
```

• `subjclit(X,Y)`

`Y` est un pronom clitique redondant sujet du verbe `X`, à droite de `X` dans une tournure interrogative (25) ou dans une phrase commençant par *peut-être, sans doute*, etc. (26) [20]. Cette tournure est caractérisée par le fait qu'une expression sujet de `X` existe à gauche de `X`. Dans ce cas, le sujet à gauche est relié au verbe par la relation `subj` et le pronom clitique à droite par la relation `subjclit`.

(25) Jacques dort-il ?
```
subj(dort,Jacques)
subjclit(dort,-il)
```

(26) Peut-être Jacques dort-il ?
```
subj(dort,Jacques)
subjclit(dort,-il)
```

Le verbe et ses compléments

Trois relations rendent compte du lien entre un verbe et l'un de ses compléments, la relation `varg`, à deux arguments, la relation `vmod` à deux arguments et la relation `vmod` à trois arguments. Le nombre d'arguments est partie intégrante de la définition de la relation et explique que deux relations ayant même nom (`vmod`) soient en fait distinguées.

La distinction entre la relation `varg` et les relations `vmod` correspond à celle qu'on fait habituellement, parmi les compléments du verbe, entre les « arguments » et les « modifieurs ». Globalement, elle correspond à la distinction entre compléments « essentiels » et « non essentiels » chez Grevisse [37, §272] :

> Les compléments [du verbe] sont essentiels [21] : 1) quand leur construction (présence ou non d'une préposition, choix de la préposition) dépend du verbe lui-même ; — 2) quand le verbe ne peut constituer sans eux le prédicat.

[19] L'information est ici clairement redondante. L'intérêt de cette redondance résidera pour nous dans le fait que l'antécédent du relatif aura la même fonction que le relatif, si bien que nous pourrons faire abstraction de ce dernier.

[20] Voir Grevisse [37, §365].

[21] Pour nous « arguments ».

Notons que, pour Grevisse, il suffit que l'une ou l'autre de ces deux conditions soit remplie pour parler de complément essentiel.

• `varg(X,Y)`

`Y` est « argument » de `X`, `X` étant le noyau d'un syntagme verbal. `Y` peut être le noyau d'un syntagme nominal (27), un pronom clitique [22] (28), le noyau d'un syntagme verbal infinitif (29), le noyau d'une proposition [23] conjonctive essentielle (30) [37, §1068] ou interrogative indirecte (31) [37, §1102], ou encore le noyau d'un syntagme adjectival complément d'un verbe à valeur de copule (p. ex. *être, sembler*) (32).

(27) Pierre boit du lait.
`varg(boit,lait)`

(28) Pierre le voit.
`varg(voit,le)`

(29) Pierre commence à boire.
`varg(commence,boire)`

(30) Il dit que Pierre aime Marie.
`varg(dit,aime)`

(31) Il demande si Pierre aime Marie.
`varg(demande,aime)`

(32) Pierre semble content.
`varg(semble,content)`

Comme pour la relation `subj`, lorsque le complément d'un verbe est une coordination de syntagmes ou de propositions, l'analyseur identifie une relation `varg` pour chacun des éléments coordonnées (33). Inversement, si plusieurs verbes sont coordonnées et partagent un même complément, l'analyseur identifie une relation `varg` pour chacun des verbes (34).

(33) Pierre mange et aime la soupe.
`varg(mange,soupe)`
`varg(aime,soupe)`

(34) Pierre boit de l'eau et du lait.
`varg(boit,eau)`
`varg(boit,lait)`

Lorsqu'il existe une relation `varg` entre un verbe et un pronom relatif, il existe aussi une relation `varg` entre ce verbe et l'antécédent du relatif (cf. la note 19 p. 256).

[22] Dans l'arbre syntaxique construit pour l'analyse du français, les pronoms clitiques compléments font partie de nœuds `FV` sans être dominés par un nœud `NP`.

[23] Une proposition est représentée par son noyau verbal.

(35) L'homme qu'il a vu n'est pas Pierre.
`varg(vu,qu')`
`varg(vu,homme)`

Enfin, notons que le complément d'agent d'un verbe à la forme passive (p. ex. *par le chat* dans *la souris a été mangée par le chat*) est considéré comme un argument du verbe [24].

• `vmod(X,Y)`

`Y` est un complément de `X` qui n'est pas argument avec les limitations suivantes. Dans la relation `vmod` à deux arguments, `Y` peut être un verbe à l'infinitif (36), le verbe noyau d'une proposition qui n'est pas argument de `X`, c'est-à-dire une proposition circonstancielle (37). `Y` peut aussi être un adverbe (38), ou le noyau d'un syntagme nominal non introduit par une préposition et à valeur adverbiale (39), ou encore un pronom disjoint à gauche du verbe (40).

(36) Pierre vient pour parler affaires.
`vmod(vient,parler)`

(37) Il boit quand il a soif.
`vmod(boit,a)`

(38) Il parle beaucoup.
`vmod(parle,beaucoup)`

(39) Il ne travaille pas le dimanche.
`vmod(travaille,dimanche)`

(40) Factofrance Heller a vu son résultat régresser de 1,4 %, dans un volume d'activités qui a été, lui, en progression de 22,1 %.
`vmod(été,lui)`

• `vmod(X,Y,Z)`

`Z` est le noyau d'un syntagme prépositionnel complément modifieur de `X`. `Y` représente la préposition qui introduit le syntagme prépositionnel dont `Z` est le noyau.

(41) D'après lui, Pierre ne viendra pas.
`vmod(viendra,d'après,lui)`

(42) Les vaches broutent dans les champs.
`vmod(broutent,dans,champs)`

[24] Grevisse dit de ce type de complément qu'il n'est « ni essentiel ni adverbial » [37, §312].

Le nom et ses compléments

Cinq relations rendent compte de la relation entre un nom et ses différents types de compléments. Deux relations (`narg`) relient un nom et un argument de ce nom ; trois relient un nom et un modifieur de ce nom (deux relations `nmod` et la relation `nn`).

• `narg(X,Y,Z)`

`Z` est le noyau d'un syntagme prépositionnel argument du nom `X`. `Y` représente la préposition qui introduit le syntagme complément.

(43) La privatisation du GAN n'aura pas provoqué de désistements.
`narg(privatisation,du,GAN)`

• `narg(X,Y)`

`Y` est le noyau d'une proposition conjonctive (44) ou le noyau d'un syntagme verbal infinitif (45) argument du nom `X`.

(44) La certitude qu'il viendra l'inquiète.
`narg(certitude,viendra)`

(45) L'envie de travailler lui manque.
`narg(envie,travailler)`

• `nmod(X,Y)`

`Y` est « modifieur » du nom `X` et n'est pas introduit par une préposition. `Y` est le plus souvent le noyau d'un syntagme adjectival ou un participe (46). Les syntagmes adjectivaux reliés à un nom sont toujours considérés comme modifieurs. `Y` peut également être un syntagme nominal en apposition, séparé du nom dont il est complément par une virgule (47), ou le verbe noyau d'une proposition relative.

(46) Le barème fiscal allemand est supérieur au français.
`nmod(barème,fiscal)`
`nmod(barème,allemand)`

(47) Maurice Lippens, le président de Fortis AG, est revenu sur sa décision.
`nmod(Maurice Lippens,président)`

• `nmod(X,Y,Z)`

`Z` est le noyau d'un syntagme prépositionnel modifieur du nom `X`. `Y` représente la préposition qui introduit le syntagme complément.

(48) La voiture de la voisine est rouge.
`nmod(voiture,de,voisine)`

• `nn(X,Y)`

`Y` est noyau d'un syntagme nominal noyau apposé sans virgule à droite de `X` et n'est pas introduit par une préposition (voir Grevisse [37, §334-335]).

(49) Le président Chirac a souscrit une assurance vie début juin.
`nn(président,Chirac)`
`nn(assurance,vie)`
`nn(début,juin)`

(50) Pour les sites les plus importants, cela se traduirait par une perte.
`nn(sites,importants)`

L'adjectif et ses compléments

Les relations entre un adjectif et un complément de cet adjectif sont globalement parallèles à celles qui sont utilisées pour les verbes. Nous ferons très peu usage de l'information selon laquelle une expression est complément d'un adjectif. Nous nous contentons donc de décrire la seule relation qui nous intéressera, `adjarg`.

• `adjarg(X,Y)`

`Y` est le noyau d'un syntagme ou d'une proposition argument de `X`. Les cas qui nous intéresseront sont ceux où `Y` est le noyau d'une proposition.

(51) Il est possible qu'il vienne.
`adjarg(possible,vienne)`

Coordination

• `coorditems(X,Y,Z)`

`X` et `Z` sont deux noyaux de syntagmes ou de propositions coordonnés. `Y` est l'expression qui exprime la coordination. Lorsque plus de deux éléments sont coordonnés, mettons une suite de la forme

e_1 et e_2 et e_3 et ... et e_n

on a $n-1$ relations `coorditems`, chacune ayant pour premier argument un élément e_i différent et qui n'est pas le dernier de la suite d'éléments coordonnés, pour troisième argument l'élément e_{i+1} de la suite et pour deuxième argument l'élément exprimant la coordination qui se trouve entre e_i et e_{i+1}. L'exemple (52) illustre ce cas.

(52) Pierre, Jacques et Jean parlent.
`coorditems(Pierre,,,Jacques)`
`coorditems(Jacques,et,Pierre)`

(53) Pierre dort quand il est fatigué ou quand il s'ennuie.
`coorditems(est,ou,ennuie)`

Verbe d'une proposition subordonnée et expression qui introduit la proposition

• `connect(X,Y)`

`X` est le noyau d'une proposition subordonnée et `Y` est une conjonction ou un pronom relatif qui introduit cette proposition.

(54) L'homme qu'il a vu n'est pas Pierre.
`connect(vu,qu')`

Noyau de syntagme nominal et déterminant

• `determ(X,Y)`

`Y` est le noyau d'un syntagme nominal et `X` est déterminant de ce syntagme nominal. `X` peut être un numéral, une des prépositions *des, du, au, aux* ou une préposition complexe se terminant par une de ces quatre prépositions (voir ci-dessus p. 244).

(55) Le chat boit du lait.
`determ(Le,chat)`
`determ(du,lait)`

Noyau de syntagme prépositionnel et préposition

• `prepobj(X,Y)`

`Y` est le noyau d'un syntagme prépositionnel et `X` est la préposition qui introduit ce syntagme.

(56) Dans ces conditions, le prix du transfert par Cera de sa participation dans le MRBB (15,1 %) est symbolique.
`prepobj(Dans,conditions)`
`prepobj(du,transfert)`
`prepobj(par,Cera)`
`prepobj(de,participation)`
`prepobj(dans,MRBB)`

Pronom relatif et antécédent

- `antec(X,Y)`

  `Y` est un pronom relatif et `X` est son antécédent.

  (57) L'homme qui rit pleurera.
  `antec(homme,qui)`

7.4.3 Traits associés aux relations

Des traits peuvent être associés aux relations comme aux nœuds de l'arbre syntaxique. Ils permettent de spécifier une sous-classe de relations à l'intérieur d'une classe de relations. Quatre sous-classes de relations seront utilisées par notre système de résolution des pronoms.

- `subj[imperso:+](X,Y)`

  `Y` est le sujet impersonnel du verbe `X`.

  (58) Il faudrait qu'il pleuve.
  `subj[imperso:+](faudrait,Il)`
  `subj[imperso:+](pleuve,il)`

- `subj[right:+](X,Y)`

  `Y` est le sujet du verbe `X` et apparaît à droite de ce verbe.

  (59) Le Crédit Agricole enverra un dossier à Bruxelles, comme le lui autorise la réglementation.
  `subj[right:+](autorise,réglementation)`

- `varg[imperso:+](X,Y)`

  `Y` est un pronom clitique accusatif complément du verbe `X` et ce pronom soit n'est pas anaphorique (60), soit renvoie à une description ou une phrase (61). Ces cas de figures sont ceux où la résolution du pronom clitique est exclue de notre objectif (voir page 173).

  (60) Le diktat du commissaire européen l'a emporté.
  `varg[imperso:+](emporté,l')`

  (61) Chacun le sait, l'hôtel Matignon use terriblement ses locataires.
  `varg[imperso:+](sait,le)`

- `nmod[appos:+](X,Y)`

  `Y` est modifieur du nom `X`, apposé à ce nom, dont il est séparé par une virgule.

(62) Maurice Lippens, le président de Fortis AG, évoque le « formidable défi » du rapprochement de Fortis avec la Générale de Banque.
`nmod[appos:+](Maurice Lippens,président)`

(63) Le président de l'AFB, Michel Freyche, a indiqué hier que les négociations reprendront début juin.
`nmod[appos:+](président,Michel Freyche)`

7.5 Apports personnels

Comme nous l'avons signalé en introduction du présent chapitre, le système qui produit l'analyse syntaxique décrite ici n'a pas été défini par nous-mêmes. Nous avons cependant été amenés à compléter le système d'analyse existant pour notre système de résolution, si bien que certains éléments décrits dans ce chapitre résultent plus particulièrement de notre propre travail. Ces éléments sont les suivants :

- passage d'une analyse phrase à phrase à une analyse globale du texte organisé comme une succession de nœuds `ST` (voir section 7.1.3) ;
- amélioration de l'identification des noms propres véritables et identification des noms propres compositionnels (voir section 7.3.1) ;
- intégration d'information à valeur sémantique pour les noms (voir p. 237 et suivantes) et les verbes de discours (voir p. 248) ;
- amélioration de l'identification des propositions incises, en particulier grâce à l'identification des verbes de discours (voir p. 248) ;
- identification des insertions entre virgules (voir section 7.3.9) ;
- amélioration de l'identification des pronoms sujet impersonnels et pronoms clitiques accusatifs non anaphoriques ou renvoyant à une description ou une phrase (non documenté).

À ces divers apports s'ajoute l'identification de bon nombre d'erreurs ou incomplétudes dans la description linguistique existante pour le français, fruit d'une observation intensive des résultats produits par le système d'analyse syntaxique défini pour le français.

Enfin, la documentation de l'analyse syntaxique constituée par le présent chapitre est également entièrement nôtre. Nous espérons avoir évité erreurs et oublis dans cette documentation. Si cet espoir est justifié, le lecteur devrait disposer, avec le présent chapitre et le suivant, de toute l'information nécessaire à l'interprétation de notre système de résolution des pronoms tel qu'il sera décrit aux chapitres 9, 10 et 11.

Chapitre 8

Formalisme

Le présent chapitre présente le formalisme dans lequel nos hypothèses sur l'interprétation des expressions pronominales seront formulées. Ce formalisme est celui du système XIP (Xerox Incremental Parser) [3] développé au Centre de recherche européen de Xerox [1].

La présentation du formalisme XIP faite ici se limite aux aspects pertinents pour que le lecteur puisse interpréter les règles que nous avons définies et qui seront présentées dans les chapitres suivants. On présente d'abord, la structure générale des règles (section 8.1), puis les différentes composantes de ces règles : les expressions régulières et la manière dont on contrôle leur instanciation (sections 8.2 et 8.3), les différents types de conditions qu'on peut poser sur les relations et les nœuds de l'arbre syntaxique (section 8.4), et enfin les différents types de conclusions possibles (section 8.5).

8.1 Structure des règles

Les règles se décomposent en trois parties :

- une expression régulière,
- un ensemble de conditions portant sur les relations entre les nœuds de l'arbre syntaxique ou les nœuds eux-mêmes, indépendamment de la structure de l'arbre,
- une conclusion.

Syntaxiquement, les règles prennent la forme suivante : l'expression régulière est placée entre deux barres verticales (« `|` ») ; elle est suivie du mot clé `if`, qui introduit les conditions, placées entre parenthèses ; les conditions sont suivies de la conclusion. Notons que soit l'expression régulière, soit les conditions, peuvent être omises. La structure des règles est présentée figure 8.1 page suivante.

[1] Nous n'avons eu aucune part dans la définition de ce formalisme. Notre travail se limite à la documentation constituée par le présent chapitre.

```
| <expression régulière> |
if ( <conditions> )
<conclusion>
```

FIG. 8.1 – Structure générale des règles.

8.2 Expressions régulières

On suppose le lecteur familier avec les expressions régulières telles qu'on les utilise le plus souvent pour analyser des suites de caractères dans des systèmes tels que Unix ou le langage de programmation Perl [2]. Dans de tels systèmes, une expression régulière dénote un ensemble de suites de caractères. Dans XIP, les expressions régulières utilisées sont un peu particulières puisqu'elles dénotent des ensembles de suites (ou « séquences ») de nœuds dans un arbre.

8.2.1 Séquences de nœuds

On appréhende les séquences de nœuds dans le contexte d'un arbre : dans un arbre donné, une séquence de nœuds est une suite de 1 à n nœuds dominés immédiatement par un même nœud.

Considérons, par exemple, l'arbre de la figure 8.2 donné à la fois sous forme graphique et sous forme parenthésée. Cet arbre contient, entre autres, les séquences de nœuds suivantes :

- `SC NP SENT`
- `NP SENT`
- `NP FV`
- `PRON VERB`

Parmi les séquences de nœuds que cet arbre ne contient pas, citons, par exemple, la séquence `FV DET` : les nœuds `FV` et `DET` dans l'arbre de la figure 8.2 ne sont pas dominés immédiatement par le même nœud.

8.2.2 Expressions simples

Les expressions régulières sont construites sur un alphabet de base qui consiste en un alphabet d'étiquettes de nœuds. Nous utiliserons ici l'alphabet des étiquettes de nœuds présenté au chapitre 7 (voir pages 222 et 225).

L'expression régulière suivante [3] :

```
| SC |
```

[2] Pour une documentation des expressions régulières, voir le site du Centre de recherche européen de Xerox, page `http://www.xrce.xerox.com/research/mltt/fst/`.

[3] Nous adoptons pour les exemples la syntaxe des formules XIP, qui veut qu'une expression régulière soit placée entre deux barres verticales.

```
                ST
        ---------------------
        |          |        |
        SC         NP      SENT
     -------       |        |
     |     |       |        |
     NP    FV    ------     |
     |     |     |    |     |
    PRON  VERB  DET  NOUN   |
     |     |     |    |     |
     Il  ferme  la  marche  .

ST{SC{NP{PRON{Il}} FV{VERB{ferme}}}
   NP{DET{la} NOUN{marche}}
   SENT{.}}
```

FIG. 8.2 – Exemple d'arbre

dénote l'ensemble des séquences de nœuds composées d'un nœud unique ayant l'étiquette `SC`.

Outre l'alphabet des étiquettes de nœuds, on dispose du symbole « `?` », qui représente un nœud d'étiquette quelconque. L'expression régulière suivante :

```
| ? |
```

dénote l'ensemble des séquences de nœuds composées d'un nœud unique, d'étiquette quelconque.

8.2.3 Expressions complexes

À partir des expressions simples, on peut construire des expressions régulières plus complexes en utilisant les opérateurs suivants.

CONCATÉNATION. Le symbole de concaténation est « `,` ». L'expression

```
| SC,NP |
```

dénote l'ensemble des séquences de nœuds commençant par un nœud `SC`, suivi d'un nœud `NP` et se terminant par ce dernier.

DISJONCTION. Le symbole de disjonction est « `;` ». L'expression

```
| SC;NP |
```

dénote l'ensemble des séquences de nœuds composées d'un nœud unique ayant soit l'étiquette `SC`, soit l'étiquette `NP`.

Étoile de Kleene. Placé à la suite de « ? » ou d'une étiquette de nœud, le symbole « * » est utilisé pour représenter une séquence de 0 à n nœuds quelconques (| ?* |) ou une séquence de 0 à n nœuds d'une étiquette donnée. L'expression

```
| SC* |
```

dénote l'ensemble des séquences de nœuds composées de 0 à n nœuds ayant l'étiquette SC.

Itération. Placé à la suite de « ? » ou d'une étiquette de nœud, le symbole « + » est utilisé pour représenter une séquence de 1 à n nœuds quelconques (| ?+ |) ou une séquence de 1 à n nœuds d'une étiquette donnée. L'expression

```
| SC+ |
```

dénote l'ensemble des séquences de nœuds composées de 1 à n nœuds ayant l'étiquette SC.

Optionnalité. Les parenthèses « (» et «) », placées autour de « ? » ou d'une étiquette de nœud, sont utilisées pour représenter une séquence de 0 à 1 nœud quelconque ou de 0 à 1 nœud d'une étiquette donnée. L'expression

```
| (SC) |
```

dénote l'ensemble des séquences de nœuds composées de 0 ou 1 nœud ayant l'étiquette SC.

Notons que l'usage de ces différents opérateurs est plus contraint dans les expressions régulières de XIP que dans les expressions régulières qui sont couramment utilisées : ces opérateurs ne peuvent porter que sur des expressions simples et non sur des expressions complexes. L'expression régulière

```
| (SC,NP) |
```

qui *a priori* pourrait dénoter l'ensemble des séquences de nœuds composées de 0 nœud ou bien d'un nœud SC suivi d'un nœud NP, n'est pas une expression régulière valide dans XIP.

8.2.4 Enchâssements de nœuds

Jusqu'à présent, les symboles et opérateurs décrits permettent de décrire des séquences de nœuds de la même manière qu'on décrirait une suite de caractères. Dans XIP, les expressions régulières offrent en outre la possibilité de décrire des nœuds en fonction des nœuds qu'ils dominent ou par lesquels ils sont dominés. En d'autres termes, elles permettent de décrire des enchâssements de nœuds.

Principe général

Pour représenter les enchâssements de nœuds, on utilise la même notation que celle que nous avons utilisée pour représenter les arbres sous forme parenthésée. Lorsqu'on veut décrire la séquence de nœuds dominés immédiatement par un nœud d'étiquette `ETIQ`,ce dernier est représenté par une formule de la forme :

```
ETIQ{...}
```

où ... indique la place d'une expression régulière qui décrit la séquence de nœuds dominée immédiatement par le nœud `ETIQ`. Les nœuds de cette séquence enchâssée peuvent à leur tour être décrits en profondeur.

EXEMPLES. L'expression régulière suivante dénote l'ensemble des séquences de nœud commençant par un nœud `SC`, qui domine immédiatement une séquence de nœuds commençant par un nœud `NP`, suivi d'un nœud `FV`. Le nœud `SC` est suivi d'un nœud `NP` et d'un nœud `SENT`.

```
| SC{NP,FV},NP,SENT |
```

La séquence `SC NP SENT` dans l'arbre présenté figure 8.2 instancie cette expression.

L'expression régulière suivante diffère de la précédente en ce que la séquence dominée immédiatement par le nœud `SC` commence par un nœud `PRON` et non un nœud `NP`.

```
| SC{PRON,FV},NP,SENT |
```

Aucune séquence de l'arbre de la figure 8.2 n'instancie cette expression, puisque le premier nœud dominé immédiatement par le nœud `SC` dans cet arbre n'a pas l'étiquette `PRON`, mais l'étiquette `NP`.

En revanche, on a de nouveau une séquence de l'arbre de la figure 8.2 qui instancie l'expression suivante, puisque le premier nœud dominé immédiatement par le nœud `SC` a bien l'étiquette `NP`, et puisque le premier nœud dominé immédiatement par ce nœud `NP` est bien `PRON`.

```
| SC{NP{PRON},FV},NP,SENT |
```

Opérateurs spéciaux

Tous les symboles et opérateurs décrits dans les sections 8.2.2 et 8.2.3 (disjonction, étoile de Kleene, etc.) sont disponibles à n'importe quel niveau de l'expression régulière. À ces opérateurs s'ajoutent deux opérateurs particuliers. Le premier sera dit « opérateur de dominance non bornée ». Le second est apparenté aux traits associés aux nœuds : il s'agit du trait `last:+`.

Opérateur de dominance non bornée. L'opérateur de dominance non bornée permet de décrire un nœud `X` en fonction d'un nœud `Y` qu'il domine à n'importe quel niveau et sans spécifier les différents niveaux d'enchâssement entre `X` et `Y`. Le symbole utilisé est `^`, qu'on accole à droite d'une étiquette de nœud, p. ex. `ETIQ^`. Un nœud peut instancier `ETIQ^` s'il a pour étiquette `ETIQ` ou s'il domine, immédiatement ou non, un nœud d'étiquette `ETIQ`.

L'expression régulière suivante dénote l'ensemble des séquences de nœuds telles que le premier nœud a l'étiquette `PRON` ou domine un nœud d'étiquette `PRON` et le second nœud a l'étiquette `NP`.

```
| PRON^,NP |
```

La séquence `SC NP` dans l'arbre de la figure 8.2 instancie cette expression régulière, puisque le nœud `SC` domine un nœud `PRON`.

Il est possible de formuler, entre accolades, une description d'arbre à droite du symbole `^`, par exemple :

```
| ETIQ1^{?*,ETIQ2} |
```

Cette description (« `?*,ETIQ2` » dans notre exemple) est alors interprétée :

- comme une description du sous-arbre dominé immédiatement par le nœud qui instancie `ETIQ1` si celui-ci a l'étiquette `ETIQ1` (c'est-à-dire que l'interprétation est la même que celle de `ETIQ1{?*,ETIQ2}`) ;
- ou comme une description du sous-arbre dominé immédiatement par un nœud d'étiquette `ETIQ1` qui est dominé par le nœud qui instancie `ETIQ1` [4].

L'expression régulière suivante dénote donc l'ensemble des séquences `s` de deux nœuds, telles que le premier nœud de `s` a l'étiquette `NP` ou domine un nœud d'étiquette `NP`, ledit nœud `NP` dominant immédiatement une séquence de 0 à n nœuds quelconques suivie d'un nœud `PRON`. Le second nœud de la séquence `s` a l'étiquette `NP`.

```
| NP^{?*,PRON},NP |
```

La suite `SC NP` dans l'arbre de la figure 8.2 instancie cette expression régulière, puisque le nœud `SC` domine un nœud `NP` qui lui-même domine immédiatement un nœud `PRON`, précédé de 0 nœud.

Trait `last:+`. Le trait `last:+` est un trait associé automatiquement par le système à tout nœud qui termine une séquence de nœuds dominée immédiatement

[4] Le point qui va un peu contre l'intuition ici est qu'en un sens, `ETIQ1` a une double interprétation : d'une part, cette suite fait référence à un nœud n_i qui n'a pas nécessairement pour étiquette `ETIQ1`, d'autre part, elle dit quelle étiquette doit avoir le nœud n_j qui domine immédiatement la séquence `?*,ETIQ2`, n_i et n_j pouvant aussi bien être un même nœud que deux nœuds différents.

par un nœud donné [5]. Par défaut, l'interprétation d'une expression `E` décrivant une séquence de nœuds sous un nœud `X` est interprétée comme étant équivalente à l'expression `E,?*`. Les deux expressions régulières suivantes ont ainsi la même dénotation :

```
| NP{?*,ADJ} |
| NP{?*,ADJ,?*} |
```

Elles dénotent toutes deux l'ensemble des séquences composées d'un nœud d'étiquette `NP` unique et qui domine immédiatement une séquence de nœuds contenant un nœud `ADJ`.

Pour contraindre l'interprétation de la suite `?*,ADJ` dans la première expression à l'ensemble des suites commençant par 0 à n nœuds quelconques et se *terminant* par un nœud `ADJ`, on ajoute, entre crochets à droite de l'étiquette `ADJ`, une condition qui exige que le nœud qui instancie `ADJ` ait le trait `last:+` :

```
| NP{?*,ADJ[last:+]} |
```

Cette expression est interprétée comme dénotant l'ensemble des séquences composées d'un nœud d'étiquette `NP` unique et qui domine immédiatement une séquence de nœuds composée de 0 à n nœuds quelconques suivis d'un nœud `ADJ` *final*.

8.3 Instanciation des expressions régulières

Les expressions régulières définies dans les règles XIP visent à être instanciées par les nœuds de l'arbre syntaxique construit sur le texte en entrée.

Lorsqu'une séquence de nœuds de l'arbre en entrée instancie l'expression régulière d'une règle, on cherche en général à identifier certains nœuds, cela au moyen de variables auxquelles on pourra faire référence. Ces variables sont présentées ci-après.

Est présentée ensuite la possibilité qu'offre le système XIP de restreindre l'interprétation d'une expression régulière à la plus longue ou à la plus courte séquence de l'arbre en entrée qui instancie cette expression.

8.3.1 Variables sur les nœuds de l'arbre syntaxique

D'une manière générale, l'objet des règles est de dire quelque chose sur les nœuds de l'arbre syntaxique et on sera donc amené à poser des conditions sur des nœuds pour aboutir à une conclusion sur ces mêmes nœuds, par exemple, l'existence d'une relation de tel ou tel type entre deux nœuds. Il nous faut donc associer des variables aux différents nœuds qui nous intéressent dans une expression régulière pour être en mesure d'y faire de nouveau référence dans la conclusion ou pour poser sur eux des conditions multiples.

[5] Pour une description du système de traits dans XIP, voir la section 7.2 au chapitre précédent.

Les variables sont représentées par le symbole « # » suivi d'un ou plusieurs chiffres. Les suites #1, #345 et #286190 sont des noms de variables valides.

Les variables sont toujours instanciées par des nœuds de l'arbre syntaxique. On les utilise aussi bien dans les expressions régulières que dans les conditions portant sur les relations entre nœuds. Dans une même règle, une même variable ne peut être instanciée que par un seul nœud, mais deux variables différentes peuvent être instanciées par le même nœud.

La règle de la figure 8.3 donne un court exemple d'utilisation des variables. Elle doit être interprétée comme suit : si une séquence de l'arbre syntaxique en entrée instancie l'expression régulière placée entre les barres verticales, alors il existe une relation de type subj entre le dernier nœud de la séquence que domine immédiatement le nœud qui instancie le nœud FV de l'expression régulière et le nœud qui instancie le nœud PRON de l'expression régulière.

```
| NP{PRON#1},FV{?*,?#2[last]} |
subj(#2,#1)
```

FIG. 8.3 – Exemple d'utilisation des variables

Appliquée à l'arbre syntaxique de la figure 8.2, cette règle donnerait lieu à la création d'une relation subj entre le nœud VERB et le nœud PRON, relation que l'on note en reproduisant les unités lexicales dominées par ces nœuds :

```
subj(ferme,Il)
```

Dans une expression régulière, les variables sont placées immédiatement à la suite des étiquettes de nœud. Dans le cas du symbole ?, la notation par la seule variable (p. ex. #1) est équivalente à ?#1.

8.3.2 Plus longue ou plus courte séquence

Non-déterminisme

L'instanciation des expressions régulières dans XIP est par défaut non déterministe, dans le sens où un même nœud pourra instancier plusieurs fois une même variable dans une expression régulière donnée, s'il existe plusieurs instanciations possibles de l'expression régulière.

Supposons qu'on veuille créer pour tout noyau de syntagme nominal une relation, qu'on appelle dans, entre ce noyau (premier argument) et le nœud ST qui le domine (deuxième argument). On définit la règle suivante :

```
| ST#2{?*,NP^{?*,#1[last]}} |
dans(#1,#2)
```

Appliquée à l'arbre de la figure 8.2, cette règle donnera lieu à la création de deux relations `dans`, à savoir :

```
dans(Il,Il ferme la marche .)
dans(marche,Il ferme la marche .)
```

En effet la séquence composée du nœud `ST` dans l'arbre de la figure 8.2 appartient à l'ensemble des séquences dénoté par l'expression régulière à deux titres. Le premier cas est celui où le premier `?*` est instancié par 0 nœud et `NP^` est instancié par le nœud `SC` ; la variable `#1` est alors instanciée par le nœud `PRON`. Le second cas est celui où le premier `?*` est instancié par le nœud `SC` et `NP^` est instancié par le nœud `NP` qui suit le nœud `SC` ; la variable `#1` est alors instanciée par le nœud `NOUN`.

Plus courte séquence

Supposons qu'on veuille maintenant restreindre l'instanciation de l'expression régulière précédente pour créer une relation `premier-dans` entre le noyau du premier syntagme nominal d'une phrase et le nœud `ST` qui le domine. On obtient cela en ajoutant le symbole `!` à droite du premier « `|` » délimitant l'expression régulière :

```
|! ST#2{?*,NP^{?*,?#1[last]}} |
premier-dans(#1,#2)
```

Appliquée à l'arbre de la figure 8.2, cette règle donnera lieu à la création d'une seule relation, à savoir :

```
premier-dans(Il,Il ferme la marche .)
```

L'instanciation du premier `?*` de l'expression par 0 nœud est plus courte que son instanciation par le nœud `SC` et est donc celle qui est retenue.

Plus longue séquence

Inversement, on peut vouloir restreindre l'instanciation de l'expression régulière à la plus longue séquence. On utilise alors le symbole `@`. La règle suivante

```
|@ ST#2{?*,NP^{?*,?#1[last]}} |
dernier-dans(#1,#2)
```

appliquée à l'arbre de la figure 8.2, donnera lieu à la création d'une seule relation, à savoir :

```
dernier-dans(marche,Il ferme la marche .)
```

8.4 Conditions

Les expressions régulières décrites dans les deux sections précédentes posent en elles-mêmes des conditions sur l'entrée susceptible d'instancier les règles. Dans la présente section, nous faisons référence par le terme « conditions » à des conditions qui ne sont pas exprimées au moyen d'une expression régulière [6]. Ces conditions, dans les règles XIP, sont placées à la suite de l'expression régulière. Elles sont introduites par le mot clé `if` et placées entre parenthèses.

Les différents types de conditions sont les suivants :

- conditions sur les traits associés aux nœuds de l'arbre syntaxique,
- conditions sur l'existence ou la non-existence d'une relation,
- conditions sur la précédence, l'identité ou la non-identité.

Ces différentes conditions peuvent être regroupées dans des formules complexes construites au moyen des opérateurs booléens de conjonction et disjonction, notés respectivement `&` et `|` dans XIP [7]. Le symbole de négation, que nous verrons utilisé dans différents contextes, est `~`.

8.4.1 Conditions sur les traits associés aux nœuds

Le système de traits dans XIP a été décrit au chapitre précédent (section 7.2).

Des traits sont associés aux nœuds de l'arbre syntaxique et il est possible, dans une formule XIP, de formuler des conditions sur les traits qui doivent ou non être associés à un nœud. Ces conditions sont placées entre crochets à la suite du symbole (étiquette de nœud ou variable) qui fait référence au nœud considéré.

On se contente ici de rappeler les différents types de conditions qui peuvent être formulées, au nombre de quatre :

(i) la condition `[att:val]` est satisfaite si le nœud a la valeur `val` pour l'attribut `att` (autrement dit, si le nœud a le trait `att:val`) ;
(ii) la condition `[att:~val]` est satisfaite si le nœud n'a pas la valeur `val` pour l'attribut `att` ;
(iii) la condition `[att]` est satisfaite si le nœud a une valeur, quelle qu'elle soit, pour l'attribut `att` ;
(iv) la condition `[att:~]` est satisfaite si le nœud n'a pas de valeur pour l'attribut `att`.

Dans le cas où l'attribut considéré ne peut prendre qu'une seule valeur, une condition de type (iii) est équivalente à une condition de type (i) et une condition de type (iv) est équivalente à une condition de type (ii).

[6] Nous verrons que les conditions sur les traits associés à un nœud pourront être insérée dans les expressions régulières.

[7] Le symbole « | » a donc plusieurs sens dans XIP : délimiteur d'expression régulière et « ou » logique. En contexte, il n'est jamais ambigu.

Les conditions sur les traits associés à un nœud peuvent être formulées aussi bien au niveau de l'expression régulière que des conditions sur les relations.

8.4.2 Conditions sur l'existence ou la non-existence d'une relation

Les règles permettent de poser des conditions sur l'existence ou la non-existence d'une relation entre nœuds de l'arbre syntaxique. Une condition sur l'existence ou la non-existence d'une relation s'entend toujours dans le contexte de l'analyse incrémentale du texte par le système XIP : les règles sont ordonnées et chaque règle est interprétée sur l'univers de dénotation construit par les règles qui l'ont précédée.

Une condition sur l'existence d'une relation est posée par simple référence à la relation. Par exemple, pour une relation de nom `nom` et d'arité 2, on écrit :

```
nom(#1,#2)
```

Cette condition est satisfaite s'il existe au moment de l'application de la règle une relation de type `nom` entre le nœud qui instancie `#1` et le nœud qui instancie `#2`.

Une condition sur la non-existence d'une relation est exprimée par ajout du symbole de négation `~` à gauche de la formule précédente :

```
~nom(#1,#2)
```

Cette condition est satisfaite s'il n'existe pas au moment de l'application de la règle de relation de type `nom` entre le nœud qui instancie `#1` et le nœud qui instancie `#2`.

Notons qu'une condition sur l'existence d'une relation est susceptible de donner lieu à l'instanciation des variables arguments de cette relation si celles-ci n'ont pas été instanciées par ailleurs. Ce n'est pas le cas pour une condition sur la non-existence d'une relation.

Des traits peuvent être associés aux relations (voir section 7.4.3 au chapitre précédent) et les différentes conditions sur les traits présentées plus haut sont aussi possibles au niveau des relations. La condition suivante, par exemple,

```
nom[att](#1,#2)
```

est satisfaite s'il existe au moment de l'application de la règle une relation de type `nom` entre le nœud qui instancie `#1` et le nœud qui instancie `#2` et si cette relation a une valeur pour l'attribut `att`.

Dans le cas des conditions négatives, la négation porte sur l'ensemble des conditions (c'est-à-dire existence et traits). Par exemple, la condition

```
~nom[att](#1,#2)
```

est satisfaite s'il n'existe pas entre le nœud qui instancie `#1` et le nœud qui

instancie `#2` de relation qui soit à la fois de type `nom` et qui ait une valeur pour l'attribut `att`.

8.4.3 Conditions sur la précédence, l'identité ou la non-identité

Précédence

En dehors des expressions régulières, il est possible de poser des conditions sur les relations de précédence entre deux nœuds. Ces conditions sont exprimées au moyen de l'opérateur « `<` », qui prend comme argument deux variables de nœud. Ces variables doivent avoir été instanciées par ailleurs. La condition

```
(#1 < #2)
```

est satisfaite si le nœud qui instancie `#1` précède celui qui instancie `#2`. On norera que les deux nœuds ne doivent pas nécessairement être dominés immédiatement pas le même nœud. Par contre, le rapport de précédence n'est pas pertinent pour deux nœuds tels que l'un domine l'autre.

Identité et non-identité de nœuds

Dans certains cas de disjonctions, il peut être utile de formuler une condition exigeant que deux variables différentes soient instanciées par le même nœud. L'opérateur utilisé est « `::` ». La condition

```
(#1 :: #2)
```

est satisfaite si `#1` et `#2` sont instanciées par le même nœud.

Inversement, on peut vouloir exiger que deux variables soient instanciées par deux nœuds différents. L'opérateur utilisé est « `~:` ». La condition

```
(#1 ~: #2)
```

est satisfaite si `#1` et `#2` sont instanciées par deux nœuds différents.

Nous verrons des exemples de conditions sur l'identité ou la non-identité de deux nœuds au chapitre 10.

Identité des traits associés à deux nœuds différents

On peut également poser des conditions sur le fait que deux nœuds aient tous les deux la même valeur pour un attribut donné. L'opérateur utilisé est de nouveau « `::` ». Le sens de cet opérateur est désambiguïsé par le fait qu'il y a référence aux traits au niveau des variables. La condition

```
(#1[att] :: #2[att])
```

est satisfaite si le nœud qui instancie `#1` et le nœud qui instancie `#2` ont la même valeur pour l'attribut `att` ou n'ont aucune valeur pour cet attribut.

8.5 Conclusions des règles

Comme nous l'avons dit (voir section 8.4.2), les règles dans XIP sont ordonnées. Chaque règle pose des conditions sur un univers de dénotation qui lui est donné par l'application des règles précédentes et vise, par sa conclusion, à modifier cet univers de dénotation. L'application d'une règle peut aboutir :

- à la création d'une nouvelle relation entre nœuds ;
- ou à l'effacement d'une relation existante ;
- ou à effectuer conjointement les deux opérations précédentes ;
- ou enfin à l'assignation de nouveaux traits aux nœud de l'arbre syntaxique.

Cette dernière opération d'assignation de traits peut être effectuée conjointement à l'une des trois opérations précédentes.

8.5.1 Création d'une relation

Une règle visant à créer une relation a comme conclusion un prédicat exprimant ladite relation.

Supposons que pour une phrase affirmative avec un sujet et un complément d'objet direct, nous voulions « extraire » une relation ternaire appelée `fait` entre le noyau du syntagme verbal (premier argument), le noyau du syntagme nominal sujet (deuxième argument) et le noyau du syntagme nominal complément d'objet direct (troisième argument). Par exemple, pour la phrase *Le chat mange la souris*, nous voulons extraire la relation

```
fait(mange,chat,souris)
```

Supposons par ailleurs que pour une phrase à la forme passive avec complément d'agent, nous voulions extraire également une relation `fait`, mais avec cette fois pour second argument le noyau du complément d'agent et pour troisième argument le sujet de la phrase. Par exemple, pour la phrase *La souris est mangée par le chat*, nous voulons extraire, comme précédemment, la relation

```
fait(mange,chat,souris)
```

En supposant que les règles déjà appliquées aient identifiées les relations `subj`, `varg` et `prepobj`, respectivement entre le verbe et son sujet, entre le verbe et son complément et entre la préposition d'un syntagme prépositionnel et son noyau, nous aurions besoin pour identifier nos deux relations des deux règles présentées figure 8.4 et 8.5, qui s'appliquent respectivement à la tournure active et à la tournure passive [8].

[8] Voir la section 7.4 pour la définition des relations utilisées ici. Le trait `dir:+` est associé à la relation `varg` si l'argument du verbe est complément d'objet direct, le trait `passive:+` à la relation `subj` si le verbe est à la voix passive, et le trait `form:fpar` au nœud lexical d'étiquette `PREP` qui domine l'unité lexicale *par*.

```
if ( subj[passive:~](#1,#2)
     & varg[dir](#1,#3) )
fait(#1,#2,#3)
```

FIG. 8.4 – Création d'une relation (1).

```
if ( subj[passive](#1,#2)
     & varg(#1,#3)
     & prepobj(#4,#3)
     & #4[form:fpar] )
fait(#1,#3,#2)
```

FIG. 8.5 – Création d'une relation (2).

Compte tenu de ce qui a déjà été dit, ces deux règles ne devraient pas présenter de difficulté. Voici néanmoins une glose pour la première : s'il existe une relation `subj`, n'ayant pas de valeur pour l'attribut `passive`, entre un nœud `#1` et un nœud `#2`, et s'il existe une relation `varg`, ayant une valeur pour l'attribut `dir`, entre le nœud `#1` et un nœud `#3`, alors il existe une relation `fait` entre `#1`, `#2` et `#3`.

8.5.2 Effacement d'une relation

Il est souvent commode de définir dans un premier temps une relation avec une définition relativement large, puis de restreindre la portée de cette définition en spécifiant un certain nombre de cas particuliers qui font exception. Ce mode de raisonnement se traduit dans XIP par la définition d'une (ou plusieurs) règles qui, s'appliquant, créent des instances de relation, instances de relation qui sont ensuite « effacées » par application de règles d'effacement de relations.

Les règles permettant d'effacer des relations se distinguent des règles qui créent des relations sur deux points :

- la conclusion de la règle se limite au symbole de négation « `~` » ;
- la relation qui doit être effacée, à laquelle il doit être fait référence dans les conditions, est préfixée par le symbole « `^` ».

À titre d'exemple, supposons que le système ait identifié pour chaque pronom anaphorique un ensemble d'instances d'une relation `coref`, telle que chaque instance relie le pronom (premier argument) à un antécédent possible (deuxième argument). En admettant qu'on ne veuille identifier pour un pronom donné qu'un seul antécédent, on pourra vouloir formuler des « préférences », par exemple dire que si un pronom sujet a parmi ses antécédents possibles une expression qui est elle-même sujet et une qui ne l'est pas, l'antécédent recherché est plus probablement le sujet. La règle présentée figure 8.6 exprime cette préférence.

```
if ( coref(#1[pron],#2)
     & subj(?,#1)
     & subj(?,#2)
     & ^coref(#1,#3)
     & ~subj(?,#3)
   ) ~
```

FIG. 8.6 – Effacement d'une relation.

La glose pour cette règle est : « s'il existe une relation de coréférence entre un nœud `#1` qui a une valeur pour le trait `pron` (un pronom, donc) et un nœud `#2`, si `#1` est sujet d'un verbe quelconque, si `#2` est également sujet d'un verbe quelconque, s'il existe par ailleurs une relation de coréférence entre `#1` et un nœud `#3` et si `#3` n'est pas sujet d'un verbe quelconque, alors effacer la relation `coref` entre `#1` et `#3` (relation marquée du symbole « `^` » dans la règle).

8.5.3 Création et effacement conjoints

Les deux opérations que sont la création et l'effacement d'une relation peuvent être effectuées conjointement dans une même règle. La syntaxe est la même que pour les règles d'effacement, c'est-à-dire qu'une des relations auxquelles il est fait référence dans les conditions est marquée d'un `^`, mais la conclusion est celle d'une règle de création, c'est-à-dire un prédicat exprimant la relation à créer.

Supposons qu'on ait deux relations binaires `subj` et `subjclit` pour rendre compte de la relation entre un verbe (premier argument) et son sujet (second argument), la relation `subj` décrivant le cas général et la relation `subjclit` visant à décrire la relation entre un pronom clitique sujet à droite d'un verbe $\mathtt{V}_i$ lorsque, dans une phrase interrogative, ce pronom est redondant par rapport à une expression sujet de $\mathtt{V}_i$ à gauche de ce dernier [9]. Par exemple, pour la phrase *Le chat dort-il ?*, on aimerait avoir les deux relations suivantes :

```
subj(dort,chat)
subjclit(dort,il)
```

Supposons qu'à un moment de l'analyse, le système ait extrait les deux relations suivantes :

```
subj(dort,chat)
subj(dort,il)
```

Nous voulons effacer la seconde de ces relations et en quelque sorte la « remplacer » par la relation `subjclit(dort,il)`. La règle présentée figure 8.7 exprime

[9] C'est l'analyse qui est effectivement faite pour le français. Voir section 7.4 p. 256.

cette opération. La variable #1 représente le verbe, la variable #2 le sujet à gauche et la variable #3 le pronom clitique.

```
if ( subj(#1,#2)
     & ^subj(#1,#3[pron])
     & (#1 < #3)
     & (#2 < #1)
   )
subjclit(#1,#3)
```

FIG. 8.7 – Effacement et création conjoints.

8.5.4 Assignation de traits

La dernière opération qui peut être effectuée dans les règles est l'assignation de traits aux nœuds de l'arbre syntaxique et/ou aux relations créées. La syntaxe est la suivante : entre les crochets « [» et «] » qui délimitent la zone réservée aux traits, on place une ou plusieurs formules sur l'un des deux modèles suivants :

- [att=val], pour assigner le trait att:val à une relation créée ou à tout nœud qui instanciera le symbole auquel la formule est associée ;
- [att=~], pour effacer tout trait ayant l'attribut att éventuellement associé aux nœuds qui instancieront le symbole auquel la formule est associée.

L'assignation de traits à un nœud n'est possible que dans le contexte d'une expression régulière. L'assignation de traits à une relation n'est possible que dans la conclusion d'une règle.

La règle présentée figure 8.8 donne un exemple d'assignation de traits aux nœuds de l'arbre syntaxique. Elle vise à identifier une relation de type subj entre le noyau #1 d'un syntagme nominal NP et le noyau #2 d'un syntagme verbal fléchi FV si les nœuds NP et FV sont au même niveau dans l'arbre et sont séparés par une suite de 0 à n nœuds qui ne n'ont pas de valeur pour l'attribut punct. Deux assignations de traits sont effectuées. D'une part, on assigne au nœud NP le trait fonc:fsubj, signifiant que ce nœud a fonction de sujet, d'autre part on assigne au nœud FV le trait argsubj:+, signifiant qu'un sujet a été trouvé pour ce syntagme verbal. Ces deux traits pourront servir pour poser des conditions sur les nœuds NP ou FV dans les règles suivantes, par exemple pour éviter qu'une règle identifiant une relation subj ne s'applique sur un nœud FV qui ait le trait argsubj:+ ou sur un nœud NP qui ait le trait fonc:fsubj [10].

[10] Cette méthode est effectivement utilisée dans le processus d'analyse syntaxique pour le français. Il y aurait d'autres moyens de formuler les mêmes contraintes, en particulier en posant des conditions sur les relations déjà extraites. Il y a donc une certaine redondance de l'information. Celle-ci est considérée comme un atout : elle permet souvent de simplifier les règles, telle ou telle formulation s'avérant mieux adaptée dans tel ou tel contexte.

```
| NP[fonc=fsubj]{?*,#1[last]},
  ?*[punct:~],
  FV[last,argsubj=+]{?*,#2[last]} |
subj(#2,#1)
```

FIG. 8.8 – Assignation de traits aux nœuds.

La figure 8.9 illustre le cas où on assigne un trait à une relation. La règle proposée reprend l'exemple de la figure 8.7, avec cette différence qu'au lieu de créer une relation dont le nom est `subjclit`, on crée une relation `subj` à laquelle on associe le trait `int:+`.

```
if ( subj(#1,#2)
     & ^subj(#1,#3[pron])
     & (#1 < #3)
     & (#2 < #1)
   )
subj[int=+](#1,#3)
```

FIG. 8.9 – Assignation de traits à une relation.

Après avoir lu le présent chapitre et le chapitre précédent, qui décrit l'analyse syntaxique qui nous est donnée en entrée du système de résolution, le lecteur dispose de toute l'information nécessaire à l'interprétation des formules qui seront présentées dans les chapitres 10 et 11, formules qui expriment notre système d'hypothèses sur l'interprétation des expressions pronominales.

Chapitre 9

Organisation du système de résolution

Muni des informations sur l'analyse syntaxique du texte en entrée du système présentée au chapitre 7 et du formalisme présenté au chapitre 8, nous pouvons maintenant aborder la partie la plus spécifique de notre système de résolution des pronoms. L'objet du présent chapitre est d'en donner une première vue globale, avant la description détaillée des différentes formules que nous avons implantées aux chapitres 10 et 11.

Nous présentons dans un premier temps les données produites par le système et rappellons les données de la clé et le prédicat d'évaluation global (9.1), puis on décrit l'organisation globale du système (9.2). Les différentes étapes du processus d'analyse sont ensuite décrites plus en détail (9.3), avec exemple des sorties intermédiaires obtenues à chaque étape de l'analyse.

9.1 Données

9.1.1 Sortie du système

L'objectif qu'on souhaite atteindre en sortie du système est de relier chaque occurrence d'une expression pronominale e_i appartenant à l'ensemble des expressions visées à au moins une expression e_j telle que e_i et e_j sont coréférentes et e_j n'appartient pas à l'ensemble des expressions retenues [1].

Un lien entre une expression pronominale e_i et sa source e_j telle qu'identifiée par le système sera représenté par une relation `coref` à deux arguments, dont le premier est l'expression pronominale et le second l'expression source, soit

`coref(`e_i`,`e_j`).`

[1] Cette présentation de notre objectif est simplifiée ; pour une spécification complète, voir la section 5.1

Considérons, à titre d'exemple, le texte suivant, où les expressions pronominales que doit interpréter le système sont en *italiques* :

(1) Le groupe Paribas$_1$ va céder la participation de 25 % qu'*il*$_1$ détient dans la banque d'affaires russe United Financial Group (UFG) au management de cette dernière. Cette décision, précise-*t-il*, *lui* permettra de conduire le développement de *ses* activités en Russie dans le cadre de *son* organisation mondiale par métier. Paribas$_2$ s'appuiera pour cela sur le bureau de représentation qu'*il*$_2$ a ouvert en 1966.

Une réponse correcte en sortie du système pourrait être [2] :

```
coref(il1,groupe)
coref(-t-il,groupe) [3]
coref(lui,groupe)
coref(ses,groupe)
coref(son,groupe)
coref(il2,Paribas2)
```

où le nom *groupe* représente le syntagme *Le groupe Paribas*.

À partir de l'ensemble des relations `coref` extraites par le système, on définit l'ensemble S_T (la « sortie », voir section 5.1.4) comme suit. On a l'ensemble R_s des reprises en sortie :

$$R_s = \{e_i \mid \text{il existe } \texttt{coref(}e_i\texttt{,}e_j\texttt{)} \text{ en sortie du système}\}$$

Pour chaque reprise e_i de R_s, on a l'ensemble A_{e_i} des « antécédents » de cette reprise :

$$\forall e_i \in R_s,\ A_{e_i} = \{e_j \mid \text{il existe } \texttt{coref(}e_i\texttt{,}e_j\texttt{)} \text{ en sortie du système}\}$$

L'ensemble S_T est un ensemble de couples :

$$S_T = \{(e_i, A_{e_i}) \mid e_i \in R_s \text{ et } A_{e_i} \text{ est l'ensemble des antécédents de } e_i\}$$

Pour l'exemple (1) ci-dessus, on a la sortie $S_{(1)}$ suivante :

$$\begin{aligned} S_{(1)} = \{ & (\text{il}_1,\{\text{groupe}\}), (\text{-t-il},\{\text{groupe}\}), (\text{lui},\{\text{groupe}\}) \\ & (\text{ses},\{\text{groupe}\}), (\text{son},\{\text{groupe}\}), (\text{il}_2,\{\text{Paribas}\}) \ \} \end{aligned}$$

La sortie du système et la définition du système lui-même sont telles qu'on peut postuler les propriétés suivantes pour chaque ensemble A_{e_i} :

$$\forall (e_i, A_{e_i}) \in S_T,\ A_{e_i} \neq \emptyset,\ e_i \notin A_{e_i}$$

[2] Nous avons vu (section 5.1.4) que des sorties différentes pouvaient être considérées comme également correctes. La sortie présentée ici est celle que le système visera effectivement.

[3] La suite *-t-il* est considérée comme une seule unité lexicale par l'analyseur syntaxique.

Un ensemble A_{e_i} quelconque est nécessairement non vide car il est défini à partir de ce qu'affiche effectivement le système en sortie.

La manière dont est défini notre système de résolution nous permet aussi d'affirmer qu'à tout moment du processus d'analyse :

$$\text{si } e_i \in A_{e_j}, \text{ alors } e_j \notin A_{e_i}$$

C'est-à-dire que la résolution des expressions pronominales est acyclique : une expression e_j ne peut être à la fois un antécédent possible d'une autre expression e_i et une reprise de cette même expression e_i.

9.1.2 Rappel des données de la clé

On rappelle ici brièvement les données spécifiées par la clé [4].

La clé spécifie l'ensemble K_T des chaînes de coréférence du texte. R_k est l'ensemble des reprises de la clé. Pour chaque élément CC_i de K_T, on a une partition en deux sous-ensembles :

$$\texttt{pron}_{CC_i} = \{e_i \in CC_i \mid e_i \in R_k\}$$

$$\texttt{src}_{CC_i} = \{e_i \in CC_i \mid e_i \notin R_k\}$$

9.1.3 Rappel du prédicat d'évaluation global

On rappelle le prédicat d'évaluation global défini dans la section 5.1.4.

La sortie de notre système de résolution des expressions pronominales sera parfaitement correcte si et seulement si :

$$\forall CC_i \in K_T, \forall e \in \texttt{pron}_{CC_i}, \exists A_e, (e, A_e) \in S_T, A_e \neq \emptyset \text{ et } A_e \subset \texttt{src}_{CC_i}$$

et

$$\forall (e_i, A_{e_i}) \in S_T, e_i \in R_k$$

Intuitivement, la première de ces deux conditions dit que pour toute reprise de la clé, il faut trouver une (ou plusieurs) source(s) correcte(s) et seulement une (ou des) source(s) correcte(s) ; la seconde condition dit qu'il ne faut trouver une source que pour des expressions qui sont des reprises dans la clé.

Par des assouplissements des exigences formulées dans la première condition de ce prédicat d'évaluation, nous spécifierons ci-après les résultats attendus aux différentes étapes du processus d'analyse.

[4] Pour une description détaillée, voir la section 5.1.3.

9.2 Organisation globale du système

Le système d'hypothèses sur l'interprétation des expressions pronominales que nous avons défini consiste en un ensemble de formules exprimées dans le formalisme propre à l'outil XIP développé au Centre de recherche européen de Xerox. Dans le système XIP, les différentes formules sont interprétées comme posant des conditions sur un univers de dénotation consistant en un arbre syntaxique et/ou un ensemble de relations entre les nœuds de cet arbre et comme résultant en une modification de cet univers de dénotation (voir les chapitres 7 et 8). Les formules sont ordonnées, si bien que l'univers de dénotation en fonction duquel une formule F_n est interprétée est celui qui résulte de l'application des formules $F_1, F_2, \ldots, F_{n-1}$.

Il résulte de cette propriété du système XIP que les formules que nous avons définies sont ordonnées. Cet ordre est le plus souvent pertinent (c'est-à-dire qu'un changement dans l'ordre des formules conduirait à un résultat différent), mais nous verrons qu'on pourra, localement, faire parfois abstraction de l'ordre.

Cela étant, les formules que nous avons définies sont regroupées en plusieurs ensembles en fonction du résultat qu'elles visent à obtenir et/ou de la manière dont elles visent ce résultat. L'organisation de ces ensembles de formules constitue la structure algorithmique de notre système d'interprétation des expressions pronominales.

Le tableau 9.1 page 288 décrit la structure de notre système. Le processus d'analyse est effectué en cinq étapes, chacune décrite par un sous-tableau du tableau global 9.1. À chaque étape correspond un ensemble de formules XIP. Dans chaque sous-tableau sont indiqués :

- en **gras**, le nom que nous donnons aux formules correspondantes ;
- ligne T, la tâche accomplie par le système ;
- ligne S, la manière dont le résultat est représenté en sortie ;
- ligne CE, le critère d'évaluation pour la tâche en question.

Intérêt d'une sortie intermédiaire. On remarquera que le critère d'évaluation indiqué dans le tableau 9.1 est le même pour les étapes 2, 3 et 4, modulo l'exigence que chaque ensemble A_{e_i} ait une cardinalité de 1 en sortie de l'étape 4. La raison de cette identité est que le critère d'évaluation proposé ne vise à évaluer que le caractère *correct* de la sortie d'une étape donnée et non l'intérêt de cette sortie. Les règles sur les zones d'antécédence (couplées aux règles sur les expressions dénotantes) définissent un ensemble de relations `coref`, qui est ensuite réduit par application des contraintes et préférences jusqu'à satisfaire l'exigence que chaque ensemble A_{e_i} ait une cardinalité de 1. Un aspect non pris en compte par le critère d'évaluation est que dès la constitution de l'ensemble *COREF* initial par les règles sur les zones d'antécédence, on souhaite avoir un ensemble *COREF* aussi réduit que possible. Le fait que les formules des différentes étapes caracté-

risent un ensemble REF_s ou *COREF* qui soit aussi réduit que possible — tout en étant correct — est ce qui fait l'intérêt des règles.

Les cinq étapes du processus d'analyse présentées dans le tableau 9.1 sont décrites plus avant dans la section suivante.

9.3 Étapes du processus d'analyse

9.3.1 Règles sur les expressions dénotantes

Les règles sur les expressions dénotantes visent à caractériser les propriétés des expressions sources, d'une part, et les propriétés des expressions reprises, d'autre part. Pour un texte donné, ces règles permettront d'identifier l'ensemble REF_s, ensemble des expressions dénotantes de ce texte selon le système. Le terme « expressions dénotantes » a une acception plus restreinte ici que dans la première partie de la thèse (voir la section 1.4). On entend par « expression dénotante » une expression qui a les propriétés requises pour être source d'une reprise ou être une reprise, étant donné l'ensemble des expressions pronominales que nous avons pour objectif d'interpréter. Compte tenu de nos objectifs, les expressions dénotantes sont toujours des syntagmes nominaux, à l'exception des déterminants possessifs.

Les expressions dénotantes sont caractérisées dans ces règles indépendamment de la présence ou non d'un pronom susceptible de les reprendre, pour les sources, et indépendamment de la présence effective d'une source, pour les reprises. Au cours de l'étape suivante, lorsque sera constitué un premier ensemble de relations de coréférence `coref(`e_i`,`e_j`)`, il sera exigé que tout argument d'une relation `coref` appartienne à l'ensemble des expressions dénotantes.

Le fait qu'une expression e_i soit une expression dénotante sera représenté par une relation unaire `ref(`e_i`)`.

EXEMPLE. Reprenons le texte de l'exemple (1) présenté page 284 :

(1) Le groupe Paribas$_1$ va céder la participation de 25 % qu'*il*$_1$ détient dans la banque d'affaires russe United Financial Group (UFG) au management de cette dernière. Cette décision, précise-*t-il*, *lui* permettra de conduire le développement de *ses* activités en Russie dans le cadre de *son* organisation mondiale par métier. Paribas$_2$ s'appuiera pour cela sur le bureau de représentation qu'*il*$_2$ a ouvert en 1966.

Les règles sur les expressions dénotantes que nous avons implantées conduiront, pour ce texte, à l'identification de l'ensemble REF_s page 289 [5].

[5] Les syntagmes nominaux identifiés comme des expressions référentielles par le système sont représentés par leur noyau.

1 - Règles sur les expressions dénotantes	
T	Caractériser un ensemble REF_s des expressions qui peuvent être source ou reprise.
S	Un ensemble de relations unaires `ref(`e_i`)` où $e_i \in REF_s$.
CE	$\forall e_i \in R_k, e_i \in REF_s$ et $\forall e_i \in R_k, \exists e_j \in REF_s, e_j \neq e_i$ et $e_j \in CC_{e_i}$

2 - Règles sur les zones d'antécédence	
T	Pour chaque reprise e_i identifiée, caractériser un premier ensemble A_{e_i} d'antécédents possibles. Par définition, un ensemble A_{e_i} quelconque est inclus dans REF_s.
S	Un ensemble *COREF* de relations binaires `coref(`e_i`,`e_j`)` où e_i est une reprise et e_j un antécédent possible de e_i.
CE	$\forall e_i \in R_s, A_{e_i} \cap CC_{e_i} \neq \emptyset$

3 - Contraintes	
T	Pour chaque couple (e_i, e_j) tel que $e_j \in A_{e_i}$, soustraire e_j de A_{e_i} s'il n'est pas « compatible » avec e_i.
S	L'ensemble *COREF* défini en 2 moins certains éléments.
CE	$\forall e_i \in R_s, A_{e_i} \cap CC_{e_i} \neq \emptyset$

4 - Préférences	
T	Pour chaque paire de couples $< (e_i, e_j), (e_i, e_k) >$, telle que $e_j \in A_{e_i}$ et $e_k \in A_{e_i}$, soustraire e_k de A_{e_i} s'il est un antécédent de e_i moins probable que e_j.
S	L'ensemble *COREF* défini en 3 moins certains éléments. Pour chaque e_i qui est une reprise en sortie, on a une seule relation `coref(`e_i`,`e_j`)`.
CE	$\forall e_i \in R_s, A_{e_i} \cap CC_{e_i} \neq \emptyset$ et $\lvert A_{e_i} \rvert = 1$

5 - Transitivité de l'antécédent vers la source	
T	À partir de l'ensemble *COREF* obtenu en 4, définir, en utilisant la transitivité de la relation `coref`, un nouvel ensemble *COREF* tel que chaque reprise soit reliée à une expressions source plutôt qu'à un antécédent qui soit lui-même une reprise
S	Le nouvel ensemble *COREF* en question. Pour chaque e_i qui est une reprise en sortie, on a une seule relation `coref(`e_i`,`e_j`)` et $e_j \notin R_s$.
CE	$\forall e_i \in R_s, A_{e_i} \cap \mathtt{src}_{CC_{e_i}} \neq \emptyset$ et $\lvert A_{e_i} \rvert = 1$

TAB. 9.1 – Organisation du système de résolution.

REF_s = {groupe, participation, il$_1$, banque, management, dernière, décision, -t-il, lui, développement, ses, activités, Russie, son, organisation, Paribas$_2$, bureau, il$_2$ }

Seules les expressions qui appartiennent à cet ensemble pourront être des reprises ou antécédents d'une reprise. Ainsi, les règles sur les expressions dénotantes disent que les syntagmes *25 %*, *d'affaires* ou *cela*, par exemple, ne peuvent être source d'une reprise par une expression pronominale appartenant à l'ensemble des expressions que nous voulons traiter.

9.3.2 Règles sur les zones d'antécédence

On suppose que l'ensemble REF_s des expressions dénotantes caractérisé par les règles sur les expressions dénotantes contient le sous-ensemble R_k, ensemble des reprises de la clé. Pour chaque élément e_i de R_k, les règles sur les zones d'antécédence visent à définir l'ensemble A_{e_i} des expressions dénotantes qui peuvent être antécédent de l'expression e_i, en vertu d'une information qui a trait essentiellement à la linéarité du texte (c'est-à-dire abstraction faite des contraintes qui seront définies par la suite). Il s'agit donc, étant donné une expression pronominale e_i dans un contexte C, de déterminer la portion de texte dans laquelle devrait se trouver une expression dénotante avec laquelle e_i soit coréférente.

On fait, avec les règles sur les zones d'antécédence, l'hypothèse que les expressions pronominales sont utilisées de telle manière que, étant donné une expression pronominale e_i, il existe dans le contexte proche de e_i, généralement avant e_i, une (voire plusieurs) expressions avec laquelle e_i est coréférente. Parmi les expressions avec lesquelles, dans un texte, une expression pronominale e_i est coréférente, on appelle « antécédent » de e_i celle(s) qui apparaisse(nt) à proximité de e_i. Nous laissons ici volontairement à la notion de proximité un contenu assez flou ; dans une certaine mesure, ce sont les règles sur les zones d'antécédence qui visent à spécifier cette notion.

Notons qu'on donne au terme « antécédent » une acception différente de « source ». Un antécédent d'une expression e_i se distingue d'une source de l'expression e_i par le fait qu'il peut lui-même être une reprise.

Là où les règles identifiant les expressions dénotantes portaient sur les expressions en elles-mêmes, les règles définissant la zone d'antécédence des expressions pronominales qui sont des reprises visent à caractériser des couples d'expressions. Pour chaque expression reprise e_i, on cherche à identifier un ensemble de couples (e_i, e_j) tels que e_j est un antécédent possible de e_i. Un couple (e_i, e_j) est représenté par une relation `coref(`e_i`,`e_j`)`.

Exemple. Pour l'exemple (1), et étant donné l'ensemble REF_s caractérisé par les règles sur les expressions référentielles, les règles sur les zones d'antécédence que nous avons implantées caractérisent un premier ensemble $COREF^Z_{(1)}$ de

relations `coref(`e_i`,`e_j`)`, à partir duquel on obtient la sortie intermédiaire $SZ_{(1)}$ suivante [6] :

$SZ_{(1)}$ = { (il$_1$,{groupe, participation}), (-t-il,{groupe, il$_1$}),
(lui,{groupe, participation, il$_1$, banque, management,
dernière, décision, -t-il}),
(ses,{décision, -t-il, lui, développement}),
(son,{décision, -t-il, lui, développement, ses, activités, Russie}),
(il$_2$,{Paribas$_2$, bureau}) }

Quelques remarques sur cette sortie intermédiaire. En premier lieu, rappelons que les couples (e_i, A_{e_i}) de l'ensemble $SZ_{(1)}$ sont déduits des relations `coref` extraites par le système. C'est parce qu'on a, par exemple, en sortie les relations :

`coref(`il$_1$`,groupe)`
`coref(`il$_1$`,participation)`

que le couple (il$_1$,{groupe, participation}) est élément de $SZ_{(1)}$ (voir la section 9.1.1).

Pour les expressions pronominales *il*$_1$, *ses, son* et *il*$_2$, les antécédents possibles en vertu des règles sur les zones d'antécédence sont les expressions dénotantes qui les précèdent dans la même phrase, pour le pronom *lui*, les expressions dénotantes de la phrase précédente et qui précèdent le pronom dans la même phrase, et pour le pronom *-t-il* les expressions dénotantes ayant fonction de sujet dans la phrase précédente et qui sont susceptibles de dénoter une personne. Cela donne une idée de la sémantique des règles sur les zones d'antécédence.

Notons que les expressions pronominales sont interprétées dans notre système relativement à des expressions et non relativement à des référents. On veut dire par là que l'information selon laquelle, par exemple, *lui* ou *-t-il* ont la même dénotation que *il*$_1$ ou *le groupe* n'est pas connue lorsqu'il s'agit d'interpréter les possessifs *ses* et *son*. Les différentes étapes du processus d'analyse sont effectuées tour à tour pour le texte entier et non phrase par phrase.

Enfin, on remarquera que cette sortie intermédiaire est parfaitement correcte au regard du critère d'évaluation présenté sur le tableau 9.1. En effet, pour chaque reprise e_i, on a bien un ensemble A_{e_i} tel que cet ensemble contient une expression e_j différente de e_i et appartenant à la même chaîne de coréférence que e_i. À ce stade de l'analyse, on ne cherche pas à relier chaque reprise à une expression qui ne soit pas une reprise (ce qui est notre objectif final).

9.3.3 Contraintes

Les contraintes sont des règles qui expriment l'impossibilité d'un lien de coréférence pour un couple (e_i, e_j) défini par les règles sur les expressions dénotantes

[6] Sortie intermédiaire qu'on peut comparer à la sortie finale $S_{(1)}$ présentée page 284.

et les règles sur les zones d'antécédence. Elles visent à réduire l'ensemble *COREF* défini par les étapes précédentes, et par là chaque ensemble A_{e_i} de l'ensemble de couples en sortie intermédiaire.

Ce sont les contraintes qui exprimeront, par exemple, la nécessité qu'une expression pronominale s'accorde en genre ou en nombre avec son antécédent.

Le terme « contrainte » est ici employé avec une acception réduite. Il est réservé ici à des règles qui aboutissent à des conclusions négatives. Les règles sur les expressions dénotantes disent « telle ou telle expression peut être une source ou une reprise ». Les règles sur les zones d'antécédence disent « tel ou tel couple (e_i, e_j) peut être un couple d'expressions coréférentes. Les contraintes disent « tel ou tel couple (e_i, e_j) *ne peut pas* être un couple d'expressions coréférentes ».

EXEMPLE. Pour l'exemple (1), nous avons en sortie des règles sur les zones d'antécédence un ensemble $COREF^Z_{(1)}$ de relations `coref(`e_i`,`e_j`)` définissant la sortie $SZ_{(1)}$ (voir page 289). Les contraintes aboutissent à la définition d'un nouvelle ensemble $COREF^C_{(1)}$ inclus dans $COREF^Z_{(1)}$. À partir de cet ensemble, on obtient la sortie intermédiaire $SC_{(1)}$ suivante :

$SC_{(1)}$ = { (il$_1$,{groupe}), (-t-il,{groupe, il$_1$}),
(lui,{groupe, participation, il$_1$, banque, management, dernière}),
(ses,{décision, lui}),
(son,{décision, lui, développement, ses, Russie}),
(il$_2$,{Paribas$_2$}) }

Huit couples (e_i, e_j) ont été soustraits de l'ensemble $COREF^Z_{(1)}$. Les quatre couples suivants l'ont été pour des raisons que nous qualifions de « structurales » :

`coref(il`$_1$`,participation)`
`coref(lui,décision)`
`coref(ses,développement)`
`coref(il`$_2$`,bureau)`

Le syntagme *la participation* est antécédent du relatif qui introduit la proposition dans laquelle *il*$_1$ est sujet. Il en est de même pour *le bureau* par rapport à *il*$_2$. Les expressions *lui* et *cette décision* dépendent du même verbe. Le possessif *ses* détermine un complément de *le développement*. Pour toutes ces raisons, les expressions considérées ne peuvent être coréférentes [7].

Le couple `coref(son,activités)` est exclu en vertu d'une contrainte qui exige que les deux expressions s'accordent en nombre.

Les trois couples suivants sont exclus car le pronom *-t-il* apparaît dans ce qu'on a appelé une « insertion » (voir p. 250), alors que les reprises potentielles

[7] Il y a en outre, pour la paire (*il*$_1$,*participation*), une incompatibilité de genre entre les deux expressions.

sont en dehors de cette insertion :

```
coref(lui,-t-il)
coref(son,-t-il)
coref(ses,-t-il)
```

Notons que dans ce dernier cas, il n'est pas correct de dire que les expressions en question ne sont pas coréférentes. L'exclusion de ces trois couples doit plutôt être interprétée comme signifiant que ce n'est pas en fonction de cette expression particulière que sont interprétées les trois reprises en question. De fait, la suite *précise-t-il* pourrait être supprimée sans que l'interprétation de ces trois expressions pronominales en souffre. Par ailleurs, la sortie $SC_{(1)}$, eu égard au critère d'évaluation posé dans le tableau 9.1, est parfaitement correcte : pour chaque reprise e_i, on a un ensemble A_{e_i} qui contient bien au moins une expression de la chaîne de coréférence à laquelle appartient e_i.

Cet exemple de sortie des contraintes, comparée à la sortie des règles sur les zones d'antécédence, permet également d'illustrer la notion d'intérêt de la sortie à différentes étapes de l'analyse, notion évoquée à la fin de la section 9.2, page 286. La sortie des contraintes est plus intéressante que la sortie des zones d'antécédence parce qu'elle propose en général un ensemble plus réduit d'antécédents possibles pour les diverses reprises. De fait, pour les pronoms il_1 et il_2, on n'a qu'un seul antécédent possible et celui-ci est correct. À chaque étape de l'analyse, on se rapproche du résultat final.

9.3.4 Préférences

Les préférences sont un ensemble ordonné de formules XIP qui visent à réduire progressivement l'ensemble $COREF^C$ obtenu en sortie des contraintes jusqu'à obtention d'un ensemble $COREF^P$ tel que pour chaque e_i dans un couple `coref(`e_i`,`e_j`)` de $COREF^P$, il n'existe qu'un seul couple `coref(`e_i`,`e_j`)`.

Les préférences sont similaires aux contraintes dans le sens où elles visent à réduire un ensemble de relations `coref` par effacement de certaines relations. Elles s'en distinguent cependant par le fait qu'elles ne s'appliquent que dans les cas d'ambiguïté, c'est-à-dire qu'une relation `coref(`e_i`,`e_j`)` ne pourra être effacée que s'il existe une relation `coref(`e_i`,`e_k`)` avec e_k différent de e_j.

Exemple. Étant donné l'ensemble $COREF^C_{(1)}$ obtenu en sortie des contraintes pour l'exemple (1), l'application de l'ensemble des préférences aboutira à la définition d'un ensemble $COREF^P_{(1)}$, inclus dans $COREF^C_{(1)}$. À partir de cet ensemble, on obtient la sortie intermédiaire $SP_{(1)}$ suivante :

$$SP_{(1)} = \{ \text{ (il}_1\text{,\{groupe\}), (-t-il,\{il}_1\text{\}), (lui,\{il}_1\text{\}),} \\ \text{(ses,\{lui\}), (son,\{ses\}), (il}_2\text{,\{Paribas}_2\text{\}) } \}$$

Onze couples (e_i, e_j) ont été soustraits de l'ensemble $COREF^C_{(1)}$. Nous n'entre-

rons pas dans les détails qui ont conduit à l'exclusion de ces couples. Le facteur déterminant ici pour les quatre expressions ambiguës avant application des préférences est le fait qu'on préfère sélectionner un antécédent qui est lui même une reprise. Dans le cas du possessif *son*, qui peut renvoyer à *lui* et *ses* (deux expressions qui sont des reprises), le possessif *ses* est retenu comme antécédent parce qu'il est le plus proche.

9.3.5 Transitivité des antécédents vers les sources

Étant donné les ensembles $COREF^P_{(1)}$ et $SP_{(1)}$ en sortie des préférences, la dernière étape du processus d'analyse consiste à relier chaque reprise à une expression qui ne soit pas elle-même une reprise, cela pour satisfaire notre objectif final et le critère d'évaluation qui l'accompagne (voir la section 5.1).

La dernière étape met simplement en jeu le fait que la relation de coréférence est transitive. Si on a les deux relations :

`coref(`e_i`,`e_j`)`
`coref(`e_j`,`e_k`)`

alors on peut déduire :

`coref(`e_i`,`e_k`)`

Récursivement, on peut aboutir à un ensemble de relations $COREF^F$ (ensemble de relations `coref` en sortie finale du système), tel que pour chaque e_i dans une relation `coref(`e_i`,`e_j`)`, il existe une relation `coref(`e_i`,`e_k`)` (e_k pouvant être identique à e_j) telle qu'il n'existe pas de relation `coref(`e_k`,`e_l`)` (c'est-à-dire que e_k n'est pas une reprise pour le système). À partir de cet ensemble $COREF^F$, on obtient la sortie finale $S_{(1)}$ suivante :

$S_{(1)}$ = { (il$_1$,{groupe}), (-t-il,{groupe}), (lui,{groupe})
(ses,{groupe}), (son,{groupe}), (il$_2$,{Paribas}) }

Cet ensemble est identique à celui qui est présenté page 284 et dont on a dit qu'il constituait une sortie parfaitement correcte pour le texte de l'exemple (1).

Cette dernière étape du processus d'analyse est relativement triviale et nous ne reviendrons pas sur elle dans la suite de la thèse.

9.4 Organisation de la suite de la présentation

Au terme de cette présentation de la structure globale de notre système d'interprétation des expressions pronominales, nous décrivons ici l'organisation de la suite de sa présentation. Nous introduisons d'abord une distinction entre « règles » et « préférences », puis nous exposons les choix de présentation adoptés pour la suite, en les justifiant.

9.4.1 Règles et préférences

Nous introduisons ici une organisation plus générale des différentes formules que nous avons définies. Les formules qui constituent le système, abstraction faite des formules qui exprime l'étape de transitivité vers les sources, sont regroupées en deux grands ensembles : les règles, d'une part, les préférences, d'autre part.

On appelle « règles » l'ensemble des formules constitué par :

- les règles sur les expressions dénotantes,
- les règles sur les zones d'antécédence,
- et les contraintes.

Les préférences restent l'ensemble des formules qui définissent l'étape 4 du processus d'analyse.

Les préférences se distinguent des règles (en particulier des contraintes ; voir la section 9.3.4) par le fait qu'elles s'appliquent toujours dans des cas où une ambiguïté sur l'antécédent d'une expression pronominale existe (c'est-à-dire quand on a au moins deux relations `coref(`e_i`,`e_j`)` et `coref(`e_i`,`e_k`)`, avec $e_j \neq e_k$). Les règles aboutissent quant à elles soit à une conclusion positive (construction d'un ensemble par les règles sur les expressions dénotantes et les règles sur les zones d'antécédence), soit à une conclusion négative, mais qui est obtenue par des conditions posées sur un seul couple <reprise, antécédent potentiel>, sans considération sur d'éventuelles alternatives à ce couple. Ainsi, les règles ont en quelque sorte une valeur « absolue » et les préférences une valeur « relative » [8].

9.4.2 Description détaillée des règles et préférences

La description des règles fait l'objet du chapitre 10 et la description des préférences celui du chapitre 11. Nous présenterons dans ces deux chapitres l'ensemble des règles et préférences qui ont été définies, en donnant l'ensemble des formules XIP implantées (celles-ci pouvant dans certains cas s'avérer plus claires qu'une glose en langue naturelle). La présentation suivra l'ordre d'application des règles et préférences, puisque celui-ci est un élément essentiel à leur interprétation [9]. Le choix de présenter les règles et préférences dans leurs moindres détails, plutôt que de nous limiter à une présentation plus générale, mais incomplète, du système, est motivé par le fait que l'évaluation qui suivra porte effectivement sur ce qui est exprimé par les règles et préférences dans leur ensemble et/ou séparément. Dans l'optique de cette évaluation, chaque détail de chaque règle ou préférence doit être pris en compte. Il ne s'agit pas de se restreindre à un ensemble réduit de caractéristiques générales qu'auraient ces règles et préférences.

[8] Pour plus de détails sur la distinction règles-préférences, voir le début de la section 11.1.

[9] Comme nous l'avons indiqué au chapitre 8 (section 8.4.2, p. 275), chaque formule XIP est interprétée sur l'univers construit par les formules qui l'ont précédée.

9.4.3 Justification des règles et préférences

Les règles et préférences expriment des hypothèses qui, dans certains cas, ne sont pas du tout nouvelles (les contraintes d'accord en genre et nombre d'un pronom avec son antécédent en sont un exemple), et qui, dans d'autres cas, sont moins évidentes. Confronté à des hypothèses du second type, le lecteur aura peut-être parfois l'impression d'un caractère arbitraire et aimerait que lui soit donnée une justification de l'hypothèse en question. Ce sentiment sera sans doute renforcé avec les préférences, dont nous verrons qu'elles visent à fournir une interprétation unique pour chaque expression pronominale, mais cela en l'absence d'informations qui semblent indispensables par ailleurs (voir la section 6.6 et la présentation des préférences au chapitre 11).

Les règles et préférences que nous avons définies sont justifiées d'une part par le fait qu'elles traduisent des hypothèses que nous avons pour beaucoup formulées à partir de l'observation d'un corpus [10], d'autre part par les résultats de l'évaluation qui sera menée par la suite, si ceux-ci confirment les hypothèses en question. Ces deux cas de figure relèvent en fait d'une même problématique, c'est la confrontation des hypothèses avec le réel qui justifie (ou invalide) les hypothèses.

S. Salmon-Alt, après avoir analysé quelques algorithmes d'interprétation des pronoms de la catégorie « robustes et pauvres en connaissances » [11], formule la critique suivante à l'égard de ces algorithmes [82, p. 85] :

> L'approche prônée par R. Mitkov, si on peut la considérer comme caricaturale, est pour autant symptomatique des modélisations « pauvres en connaissances » : s'il y a effectivement, à court terme, un besoin en systèmes opérationnels, baser un modèle qui se veut scientifique sur un faisceau d'heuristiques dont l'auteur ne prend même plus la peine de donner ne-serait-ce que le début d'une justification théorique (Pourquoi un indéfini a-t-il moins de chances d'être l'antécédent d'un pronom ? D'où vient la liste des verbes et des connecteurs ? Comment justifier la pondération des scores ? Pourquoi la limitation aux deux phrases précédentes ?) comporte un double risque pratique et théorique : d'un point de vue pratique, de tels systèmes sont difficilement maintenables et peu évolutifs, car ils fournissent toujours une réponse, ne signalent jamais de problèmes et sont incapables de fournir des

[10] C'est-à-dire que pour le corpus que nous avons étudié, les règles et préférences ont dans l'ensemble une certaine validité, même si celle-ci n'est quasiment jamais absolue. On aurait pu donner dans le courant de la présentation les résultats produits par les différentes règles et préférences sur le corpus d'étude, dans la mesure où cela aurait levé une certaine impression d'arbitraire, mais il nous a semblé plus pertinent de rassembler ces résultats au chapitre 12, avec ceux de l'évaluation sur un nouveau corpus.

[11] Ceux de Hobbs, Lappin & Leass et Mitkov. Voir la description de ces systèmes dans la section 6.5.

> indications de la source éventuelle des problèmes. D'un point de vue théorique, ils ne valident aucune hypothèse scientifique et ne contribuent en rien à l'explication des processus linguistiques et cognitifs sous-jacents à la résolution des anaphores.

Si nous adhérons à certaines des remarques de Salmon-Alt [12], nous ne pouvons la suivre sur la majeure partie de son discours.

En premier lieu, relevons un usage selon nous inapproprié de l'adjectif *scientifique* : pour nous, il n'y a rien de tel qu'un « modèle scientifique » ou qu'une « hypothèse scientifique », il y a des modèles ou hypothèses qui sont testables et testés dans une *démarche* scientifique. La démarche de Mitkov, qui teste effectivement ses hypothèses, *est* scientifique.

Salmon-Alt reproche à Mitkov l'absence de « justification théorique » de son système. En tant que système d'hypothèses effectivement testable, l'algorithme de Mitkov *est* une théorie [13]. Les deux *Pourquoi ?* de Salmon-Alt nous indiquent ce que serait pour elle une « justification théorique ». Ces deux questions nous semblent hors de propos : Mitkov fait les deux hypothèses correspondant à ces deux questions parce qu'il a fait certaines observations et qu'il pense avoir isolé là ce qui est peut-être un facteur déterminant dans l'interprétation des pronoms. Si tous les pronoms dont Mitkov vise à spécifier l'interprétation, dans un type de corpus dûment spécifié, étaient toujours interprétés comme coréférents avec une expression de la même phrase ou de l'une des deux phrases qui précèdent, cela ne constituerait-il pas une justification suffisante à son hypothèse ? On peut s'intéresser au *pourquoi* du problème de l'interprétation des pronoms, mais Mitkov s'intéresse clairement au *comment*. Il n'y a pas lieu de le lui reprocher, ni d'associer au premier type de question un statut privilégié par rapport au second [14].

Pour terminer, en regard de la dernière phrase de Salmon-Alt dans la citation ci-dessus, nous tenons à préciser que si l'évaluation des systèmes de Hobbs, Lappin & Leass, Mitkov (et d'autres) invalide globalement les hypothèses exprimées dans ces systèmes, ces derniers ont au moins le mérite de pouvoir être et d'avoir été évalués. Le fait que les hypothèses initiales n'aient pas été confirmées par l'évaluation n'invalide pas la démarche qui a consisté à les formuler explicitement, puis à les tester.

---

[12] Pour ce qui concerne le problème de la pondération optimale des scores et la difficulté à analyser les échecs avec un système de préférences pondérées. Voir plus loin une discussion de ces points dans la section 11.1.1.

[13] Voir la citation de Bès [11] dans la section 2.7.1, page 80.

[14] Les conditions d'évaluation de réponses éventuelles aux questions en *pourquoi* du type de celles de Salmon-Alt étant probablement assez difficiles à mettre en place, nous ajouterons qu'associer ce type de problème au terme *théorie* risque fort de donner à ce dernier le sens péjoratif de « système non évaluable en pratique ». Dans ce sens, le « point de vue théorique » de Salmon-Alt ne nous paraît pas intéressant.

Chapitre 10

Règles

La présentation des hypothèses qui constituent le cœur de notre système d'interprétation automatique des expressions pronominales est répartie sur deux chapitres. Ce chapitre présente l'ensemble des « règles », qui définissent les trois premières étapes du processus d'analyse.

Les formules présentées font usage de l'information syntaxique présentée au chapitre 7. Pour chaque élément d'information utilisé dans le présent chapitre, on indique l'endroit où est présentée cette information au chapitre 7 par une référence à une page ou sous-section de ce chapitre, référence préfixée de « as » (pour « analyse syntaxique ») et placée entre crochets. Par exemple, la suite [as-256] indique que l'information utilisée se trouve page 256.

Les formules sont exprimées dans la syntaxe du formalisme du système XIP. Celui-ci a été présenté au chapitre 8. Les différentes formules sont données littéralement dans une présentation similaire à celle des figures ou tableaux. Chaque ligne d'une formule littérale est numérotée. On indiquera la correspondance entre la description ou glose d'une formule donnée dans le corps du texte avec la formule littérale par l'indication de la ligne concernée entre parenthèses. Par exemple, la suite « (l. 5) » veut dire « voir la ligne 5 de la formule décrite ».

Les objectifs des différentes règles et leurs sorties respectives ont été décrits au chapitre précédent section (9.3). On se contente ici de les rappeler brièvement au début de chaque section.

10.1 Règles sur les expressions dénotantes

Les règles décrites dans la présente section visent à caractériser l'ensemble REF_s des expressions qui peuvent être des reprises ou antécédents de reprises. Dans la mesure où les liens de reprise que nous recherchons sont toujours des liens de coréférence (au sens défini dans la section 2.1), nous disons de ces expressions qu'elles sont des « expressions dénotantes ».

La caractérisation des expressions dénotantes constitue la première étape du processus d'interprétation des expressions pronominales. Cette caractérisation est elle-même effectuée en deux temps : on caractérise d'abord de manière générale l'ensemble des expressions dénotantes, puis on formule des exceptions sur l'ensemble défini dans un premier temps. Ces deux parties de la définition des expressions dénotantes sont présentées successivement dans les deux sous-sections suivantes.

Dans le système tel qu'il est implanté, on représente le fait qu'une expression est une expression dénotante en créant une relation unaire `ref` (pour « référent ») dont l'argument est cette expression. Dans le cas des syntagmes nominaux, l'argument de la relation `ref` n'est pas le syntagme complet, mais seulement son noyau, qui est considéré comme représentant du syntagme complet. Le premier temps de la définition consiste en un ensemble de formules qui aboutissent à la création de relations `ref` ; le second temps consiste à effacer certaines relations de l'ensemble des relations `ref` définies.

Une même expression peut tout à fait satisfaire les conditions posées par plusieurs règles. Les règles définissant les expressions dénotantes doivent être vues comme une conjonction, c'est à dire qu'elles sont non ordonnées, si on excepte l'ordre qui veut que la définition générale soit posée avant les exceptions.

Les règles sur les expressions dénotantes seront référencées par la suite ER suivie d'un numéro d'ordre.

10.1.1 Définition générale

La définition générale des expressions dénotantes spécifie un ensemble sur lequel nous formulerons par la suite des restrictions. Pour cette raison, nous disons des règles présentées dans la section présente qu'elles spécifient des expressions dénotantes « potentielles », une expression dénotante étant au final une expression qui satisfait les conditions de la définition générale et ne satisfait pas les conditions spécifiées par les restrictions posées dans la section suivante (section 10.1.2).

ER.1 - Syntagmes nominaux déterminés. Un syntagme nominal déterminé est une expression dénotante potentielle. Cette règle est simple : s'il existe une relation `determ` entre une expression `#1` et une expression `#2`, alors il existe `ref(#2)`. La relation `determ` relie le déterminant d'un syntagme nominal (premier argument) au noyau dudit syntagme (deuxième argument) [as-261].

```
1  if ( determ(#1,#2) )
2  ref(#2)
```

Règle ER.1 – Syntagmes nominaux avec déterminant.

ER.2 - Noms propres, sigles, numéraux. Un syntagme nominal dont le noyau est un nom propre [as-234], un titre [as-234], ou un numéral [as-242] est

une expression dénotante potentielle. La règle met en jeu à la fois une expression régulière et des conditions sur les traits associés au nœud qui instancie la variable `#1` dans l'expression régulière. Elle peut être glosée comme suit : étant donné un syntagme nominal dont le noyau est `#1`, si `#1` a une valeur pour l'attribut `proper` ou pour l'attribut `tit` ou pour l'attribut `num`, alors `#1` est une expression dénotante potentielle.

```
1  | NP{?*,#1[last]} |
2  if ( #1[proper]
3          | #1[tit]
4          | #1[num]
5     )
6  ref(#1)
```

Règle ER.2 – Noms propres, titres, numéraux.

En *italiques* et entre crochets dans les deux phrases suivantes, les expressions dénotantes caractérisées par la règle ER.2.

(1) Reste quatre administrateurs : parmi eux, [*trois*] ont démissionné de leur fonction en début d'année.

(2) « Notre priorité est de lancer avec succès ces deux nouvelles sociétés », a souligné [*M.*] [*Nishimura*].

L'exemple 1 illustre le fait que les pronoms numéraux ne sont pas caractérisés comme des pronoms en sortie de l'analyse syntaxique, mais comme des unités lexicales dominées par un nœud d'étiquette `NUM` et qui constituent à elles seules un nœud `NP`.

L'exemple 2 illustre le fait que le titre qui précède un nom propre et le nom propre lui-même sont dominés par deux nœuds `NP` distincts [as-235], la seconde expression étant considérée comme un syntagme apposé à la première [as-260]. À ce titre, elle sera exclue de l'ensemble des expressions dénotantes (voir ci-dessous la règle ER.9) et c'est l'expression *M.* qui représentera la référence à l'être de l'univers qu'est M. Nishimura.

ER.3 - Pronoms. Un pronom est une expression dénotante potentielle.

```
1  | PRON#1 |
2  ref(#1)
```

Règle ER.3 – Pronoms

ER.4- Déterminants possessifs. Un déterminant possessif est une expression dénotante potentielle. Plus précisément, la règle exige que le déterminant possessif fasse partie d'un syntagme nominal et soit suivi dans ce syntagme d'au

moins une expression. Cette exigence permet de pallier certaines erreurs d'analyse syntaxique où la suite *SA* ou *Sa* signifiant « société anonyme » dans un nom de société (p. ex. *la banque mexicaine Confia Sa*) est mal désambiguïsée.

```
1  | NP{?*,DET#1[poss],?} |
2  ref(#1)
```

RÈGLE ER.4 – Déterminants possessifs

ER.5 - PREMIER SYNTAGME NOMINAL D'UNE PHRASE SANS PROPOSITION PRINCIPALE. Lorsqu'il est le premier syntagme nominal d'une phrase sans proposition principale, un syntagme nominal est une expression dénotante potentielle. On vise surtout par cette règle des syntagmes nominaux qui ne satisfont pas les conditions des règles ER.1 et ER.2.

Si une phrase contient une proposition principale, celle-ci est toujours représentée par le premier nœud dominé immédiatement par le nœud `ST`, modulo la présence éventuelle de symboles de ponctuation ou de balises de type SGML. Les conditions sur les traits associés aux nœuds susceptibles d'instancier `?*` au début de `ST` excluent les propositions noyau (la phrase ne contient donc pas de principale) et les syntagmes nominaux ou prépositionnels.

```
1  | ST{?*[sc:~,np:~,pp:~],NP{?*,#1[last]}} |
2  ref(#1)
```

RÈGLE ER.5 – Premier syntagme nominal d'une phrase sans principale

À titre d'exemple, les deux nœuds `ST` dans l'arbre suivant, analyse de la suite la suite *Autre candidat à devoir faire ses preuves : Eureko*, instancient l'expression régulière de la règle ER.5 :

```
ST{NP{AP{ADJ{Autre}} NOUN{candidat}}
   IV{PREP{à} VERB{devoir}}
   IV{VERB{faire}}
   NP{DET{ses} NOUN{preuves}}
   PUNCT{:}}
ST{NP{NOUN{Eureko}}
   SENT{.}}
```

Deux nœuds de cet arbre instancient la variable `#1` : d'une part le nœud lexical de *candidat*, qui représentera pour nous le syntagme *Autre candidat*, d'autre part le nœud lexical de *Eureko*. Notons que cette dernière expression satisfait aussi les conditions posées dans la règle ER.2, mais que ce n'est pas le cas de la première.

ER.6 - SYNTAGMES NOMINAUX SUJETS. Lorsqu'il est sujet d'un verbe, un syntagme nominal est une expression dénotante potentielle. Comme pour la règle précédente, on vise par cette règle essentiellement des syntagmes nominaux qui

ne satisfont pas les conditions posées par les deux premières règles. L'exigence que le nœud qui instancie `#1` n'ait pas de valeur pour l'attribut `verb` exclut les sujet verbaux, comme dans la phrase *Dormir est un plaisir.*

Dans l'exemple suivant, le syntagme *Pareille fatalité* est sujet du verbe *mène* et, à ce titre, est une expression dénotante potentielle. Il existe une relation `ref` dont l'argument est l'expression *fatalité*. Notons que le syntagme *Pareille fatalité* ne satisfait aucune des conditions posées par les règles précédentes.

(3) Pareille [*fatalité*] ne mène pas seulement, entropie aidant, de l'idée pure au formalisme [...].

```
1  if ( subj(#2,#1[verb:~])
2  ref(#1)
```

RÈGLE ER.6 – Syntagmes nominaux sujets

10.1.2 Exceptions

L'ensemble d'expressions dénotantes caractérisé par notre définition générale est plus large que nous le souhaitons : parmi les expressions qu'il contient se trouvent des expressions que nous avons exclues de notre champ d'étude (p. ex. les pronoms de première ou deuxième personne), des expressions pronominales dont l'interprétation ne met pas en jeu une relation de coréférence (p. ex. les pronoms sujets impersonnels), ou encore des syntagmes nominaux dont on pense qu'ils ne pourront pas être repris par une expression pronominale. On formule donc un ensemble de règles pour exclure ces expressions de l'ensemble défini par la définition générale. Ces règles sont présentées ici d'abord pour les expressions pronominales, puis pour les syntagmes nominaux non pronominaux.

Restrictions sur les expressions pronominales

ER.7 RESTRICTIONS SUR LES EXPRESSIONS PRONOMINALES. Parmi les expressions pronominales exclues de notre champ d'étude, certaines ne pourront pas être antécédent d'une reprise par une expression pronominale retenue. La règle ER.7 exclut ces expressions qui ne peuvent être antécédent d'une reprise de l'ensemble des expressions dénotantes.

Sont exclus de l'ensemble des expressions dénotantes :

- les expressions de première ou deuxième personne [as-247] (l. 2-3) ;
- les pronoms réfléchis, relatifs, interrogatifs, clitiques « génitif » (*en*), clitique « locatif » (*y*) [as-7.3.2] (l. 4-8) ;
- les pronoms démonstratifs neutre (p. ex. *ceci, ce*) [as-243] (l. 9) ;
- les pronoms sujet impersonnels [as-262] (l. 10) ;

```
1  if ( ^ref(#1)
2       & ( #1[p1]
3         | #1[p2]
4         | #1[refl]
5         | #1[rel]
6         | #1[int]
7         | #1[clit,gen]
8         | #1[clit,loc]
9         | #1[neutre]
10        | subj[imperso](?,#1)
11        | varg[imperso](?,#1)
12        )
13      ) ~
```

RÈGLE ER.7 – Restrictions sur les expressions pronominales.

– les pronoms clitiques accusatif qui ne sont pas référentiels ou renvoient à une phrase [as-262] (l. 11).

S'il existe une relation `ref` pour ces expressions, celle-ci est effacée (voir la section 8.5.2).

Restrictions sur les syntagmes nominaux non pronominaux

Un certain nombre de syntagmes nominaux non pronominaux caractérisés comme des expressions dénotantes potentielles par la définition générale ne pourront pas selon nous être source d'une reprise par une expression pronominale coréférente. Il s'agit des syntagmes indéfinis qui ont fonction d'attribut, des syntagmes qui sont en apposition, des syntagmes qui expriment des mesures et des syntagmes qui ont fonction de complément dans des noms propres tels que *le Mouvement contre le racisme et pour l'amitié entre les peuples*. Ces expressions sont exclues de l'ensemble des expressions dénotantes potentielles par les quatre règles suivantes.

ER.8 - SYNTAGME NOMINAL INDÉFINI ATTRIBUT. On appelle syntagme nominal attribut un syntagme nominal non prépositionnel complément d'un verbe à valeur de copule (p. ex. *être, rester, paraître*) [as-247]. Un syntagme nominal attribut n'est pas une expression dénotante (c'est-à-dire qu'il ne pourra pas être repris par une expression pronominale) s'il est déterminé par un déterminant indéfini ou quantifieur (l. 3) [as-7.3.3], ou par un numéral.

Font cependant exception à cette règle les cas où le syntagme en question est attribut d'un verbe dont le sujet est un pronom démonstratif neutre [as-243]. Dans cette configuration, on considère que l'attribut, plutôt que le pronom, représente le référent, dans la mesure où nous avons exclu de l'ensemble des expressions dénotantes les pronoms démonstratifs neutres.

Considérons l'exemple suivant :

(4) Le Raroc est un bon outil de gestion du capital [...]. C'est aussi un outil de gestion du temps : sans l'utiliser comme moyen de tarification, nous essayons de hiérarchiser le temps d'investissement des commerciaux en fonction du capital économique alloué.

Le syntagme *Le Raroc*, le pronom démonstratif *C'* et le pronom *l'* sont coréférents. Cependant nous ne traitons pas les reprises par pronom démonstratif et identifier un lien de coréférence entre *l'* et *C'* serait une réponse correcte en sortie de notre système (voir section 5.1.4, page 177).

Pour avoir une réponse plus intéressante, nous avons choisi de nous donner dans ce cas pour objectif de relier le pronom *l'* à l'expression *un outil de gestion du temps*. Cette dernière expression est une expression dénotante tandis que le pronom *C'* n'en est pas une, suivant la règle ER.7. Ce choix, qui n'est pas exempt d'arbitraire, est motivé par le fait qu'on dispose, grâce à la présence de l'unité lexicale *outil*, d'une information plus riche sur le référent auquel s'applique cette description.

Cela étant, la règle est la suivante :

```
1  if ( ^ref(#2)
2       & varg(#1[copule],#2)
3       & determ(#3,#2)
4       & ( #3[indeter] | #3[num] )
5       & ~subj(#1,?[pron,dem,neutre])
6     ) ~
```

RÈGLE ER.8 – Restriction sur les syntagmes attributs.

Glose : Une expression dénotante potentielle `#2` qui est argument [as-257] d'un verbe à valeur de copule (`#1[copule]`) et est déterminée par une expression `#3` qui a soit le trait `indef:+`, soit le trait `quant:+` [1], soit le trait `num:+` n'est pas une expression dénotante, sauf si le sujet du verbe copule est un pronom démonstratif neutre (`#3[pron,dem,neutre]`).

ER.9 - SYNTAGME NOMINAL APPOSÉ. Un syntagme nominal apposé n'est pas une expression dénotante. Rappel : on appelle « syntagme nominal apposé » un syntagme nominal noyau complément d'un autre syntagme nominal noyau dans l'une ou l'autre des deux configurations suivantes :

- syntagme nominal sans préposition complément d'un autre syntagme nominal sans virgule séparatrice (p. ex. *Chirac* dans *le président Chirac*) [as-260] ;
- syntagme nominal sans préposition complément d'un autre syntagme nominal avec virgule séparatrice (p. ex. *le président* dans *Maurice Lippens, le président de Fortis AG, est revenu sur sa décision*) [as-262].

[1] Ces deux traits étant subsumés par l'attribut général `indeter` (voir figure 7.16 p. 244).

L'information selon laquelle un syntagme nominal est apposé nous est donnée par la relation `nn` dans le premier cas et `nmod` dans le second.

```
1  if ( ^ref(#1)
2       & ( nmod(#2,#1)
3           | nn(#2,#1) )
4     ) ~
```

RÈGLE ER.9 – Restriction sur les syntagmes nominaux apposés.

Dans la phrase :

(5) La ministre de la Justice Elisabeth Guigou a rappelé hier devant la commission des Lois qu'un décret circule au sein de la profession des mandataires et administrateurs liquidateurs des entreprises.

le syntagme *Elisabeth Guigou* est apposé au syntagme *La ministre de la Justice*, ce qui est représenté en sortie de l'analyseur syntaxique par la relation suivante :

```
nn(ministre,Elisabeth Guigou)
```

L'expression *Elisabeth Guigou* n'est pas une expression dénotante ; s'il y a reprise par une expression pronominale dénotant Elisabeth Guigou, alors c'est au nom *ministre* que cette expression pronominale devra être reliée.

ER.10 - MESURES. Les expressions qui expriment la mesure d'un être de l'univers de dénotation ne peuvent être source d'une reprise pronominale avec coréférence [as-240]. Exemples : *25 degrés* dans *la température est de 25 degrés* ; *20 milliards de francs* dans *Sa production annuelle s'élève à 20 milliards de francs.* Les noms d'unités de mesures sont repérés par le fait que le trait `measure:+`leur est associé. En règle générale, un syntagme nominal exprimera une mesure d'un être de l'univers de dénotation si son noyau est un nom d'unité de mesure et s'il est déterminé par un numéral, un déterminant indéfini ou un déterminant quantifieur [as-7.3.3].

```
1  if ( ^ref(#1[measure,indeter]) ) ~
```

RÈGLE ER.10 – Restriction sur les syntagmes exprimant une mesure.

ER.11 - PARTIE DE NOM PROPRE COMPOSITIONNEL. Un syntagme nominal qui est partie d'un nom propre compositionnel [as-236] ne peut être antécédent d'une reprise pronominale avec coréférence. Un tel syntagme nominal est caractérisé dans l'entrée par le fait que son noyau a le trait `pnpart:+`.

Étant donné la phrase,

(6) Pierre a adhéré au Mouvement contre le racisme et pour l'amitié entre les peuples.

```
1  if ( ^ref(#1)
2       & #1[pnpart] ) ~
```

RÈGLE ER.11 – Restriction sur les syntagmes parties de noms propres.

les syntagmes nominaux noyau *le racisme*, *l'amitié* et *les peuples* sont exclus de l'ensemble des expressions dénotantes, c'est-à-dire qu'ils ne pourront pas être repris par une expression pronominale.

ER.12 - SYNTAGMES COMPLÉMENTS D'UN NOM DE FRACTION. Un syntagme nominal complément d'un nom de fraction (p. ex. *quart, plupart*) [as-238] ou d'un pronom quantifieur (p. ex. *chacun, beaucoup*) [as-7.3.2], introduit par la préposition *de* (ou *des*) et déterminé par une forme au pluriel ne peut être repris par une expression pronominale.

```
1  | NOUN#3[fraction];PRON#3[quant] |
2  if ( nmod(#3,#2[form:fde],#1)
3       & determ(?[pl],#1)
4       & ref(#3)
5       & ^ref(#1)
6     ) ~
```

RÈGLE ER.12 – Compléments d'un nom de fraction ou pronom quantifieur.

Selon la règle ER.12, le syntagme *des greffiers* dans la phrase suivante n'est pas une expression dénotante (c'est-à-dire qu'il ne peut être repris par un pronom). L'idée essentielle derrière la règle est que s'il y a reprise, alors l'expression pronominale dénotera le sous-ensemble constitué du quart des greffiers et non l'ensemble entier des greffiers.

(7) Dans cette hypothèse, souligne Henri Nappi, « seul le quart des greffiers verrait ses ressources équilibrées ».

Notons que par un procédé que nous ne documenterons pas dans la thèse [2], les traits véhiculant l'information à valeur sémantique [as-237] associée au noyau d'un syntagme SN_i exclu des expressions dénotantes par la règle ER.12 sont transférés au noyau du syntagme dont SN_i est complément. Ainsi, dans l'exemple (7), le nœud lexical de *quart* reçoit les traits à valeur sémantique qui sont associés au nœud lexical de *greffiers*, en l'occurrence le trait `person:+` [as-240].

Cette même opération de transfert des traits à valeur sémantique est également effectuée entre les les noms à valeur de numéral [as-238] et leur « pseudo-complément » (voir [37, §422]). Le pseudo-complément en question est caractérisé comme un syntagme prépositionnel introduit par *de* et sans déterminant. Ainsi,

[2] Il met en jeu un type de formule XIP très spécifique, que nous jugeons superflu de documenter pour le seul cas particulier en question ici.

si le texte analysé contient, par exemple, le syntagme *une vingtaine de greffiers*, on aura l'analyse syntaxique suivante :

```
NP{DET{une} NOUN{vingtaine}} PP{PREP{de} NP{NOUN{greffiers}}}
nmod(vingtaine,de,greffiers)
```

Le syntagme nominal noyau *une vingtaine* est une expression dénotante, mais pas le syntagme nominal noyau *greffiers* (il ne satisfait aucune des conditions posées par les règles ER.1 à ER.6). Le syntagme nominal *une vingtaine de greffiers* est représenté par le nom *vingtaine*, auquel sont transférés les traits à valeur sémantique associés au nœud lexical de *greffiers*.

10.1.3 Relation entre chaque expression dénotante et la phrase qui la contient

Nous aurons besoin par la suite de pouvoir tester si deux expressions dénotantes figurent dans la même phrase, ou dans deux phrases différentes telles que l'une précède immédiatement l'autre, etc. L'analyseur XIP, conçu au départ pour une analyse phrase par phrase, ne fournit pas de mécanisme direct pour effectuer de tels tests. On se donne donc, par la règle ER.13, une relation `dans`, à deux arguments, qui relie une expression dénotante au nœud `ST` qui la domine :

```
1  | ST#2{?*,?^#1}} |
2  if ( ref(#1) )
3  dans(#2,#1)
```

RÈGLE ER.13 – Relation `dans`.

La règle se lit comme suit : étant donné un nœud `#2` étiqueté `ST` et `#1` un nœud situé à n'importe quel niveau sous le nœud `#2` (l. 1), si `#1` est une expression dénotante (l. 2), alors créer la relation `dans(#2,#1)` (l. 3). Le premier argument d'une relation `dans` est donc un nœud phrase et le second une expression dénotante dominée par ce nœud.

Dans la mesure où la règle ER.13 s'applique pour toute expression dénotante, l'existence d'une relation `dans(#2,#1)` implique l'existence d'une relation `ref(#1)`. Par la suite, pour poser une condition sur le caractère dénotant d'une expression `#1`, nous utiliserons la relation `ref(#1)` ou la relation `dans(#2,#1)`, selon que l'information sur la phrase où figure `#1` est ou n'est pas pertinente.

10.2 Règles sur les zones d'antécédence

Étant donné l'ensemble des expressions dénotantes potentielles Ref_s défini par les règles sur les expressions dénotantes, les règles sur les zones d'antécédence définissent un ensemble de couples (e_i, e_j), où e_i est une reprise et e_j un antécédent

possible de cette reprise en vertu d'une information qui a trait essentiellement à la linéarité du texte. Il s'agit, étant donné une expression pronominale dénotante dans un contexte C, de déterminer la portion de texte dans laquelle devrait être trouvée une expression dénotante avec laquelle e_i est coréférente.

Les couples (e_i, e_j) sont représentés en sortie par des relations `coref(`e_i`,`e_j`)`. Les règles sur les zones d'antécédence aboutissent donc toujours à la création d'une relation `coref`. Les règles sur les zones d'antécédence mettent toujours en relation des expressions dénotantes, si bien qu'une règle contient toujours une condition exigeant qu'il existe une relation `ref(`e_i`)` (ou `dans( ?,`e_i`)`) et une condition exigeant qu'il existe une relation `ref(`e_j`)` (ou `dans( ?,`e_j`)`) pour aboutir à la conclusion qu'il existe une relation `coref(`e_i`,`e_j`)`.

Les relations `coref` définies par les règles sur les zones d'antécédence expriment une première hypothèse sur l'existence d'une relation de coréférence entre les deux expressions et non des affirmations sur l'existence de ces relations. Un prédicat `coref(#1,#2)`, que l'on retrouvera en conclusion des règles définissant les zones d'antécédence, doit donc être interprété comme signifiant « l'expression `#1` peut être coréférente avec l'expression `#2`, abstraction faite des contraintes ou préférences définies par ailleurs. »

Organisation des règles. Nous distinguons, dans la définition des règles caractérisant les zones d'antécédence, trois types d'expressions pronominales :

- les pronoms clitiques,
- les pronoms disjoints,
- les déterminants possessifs.

Les trois sections suivantes présentent les règles définies pour chacun de ces trois types d'expressions, successivement. L'identification de ces règles, pour référence ultérieure, sera effectuée au moyen de sigles composés d'un « Z » initial (pour « zone »), suivi de « PC », ou « PD », ou « DP », respectivement pour « pronom clitique », « pronom disjoint » et « déterminant possessif », suivi d'un chiffre.

Pour chaque type d'expressions pronominales, les règles sont ordonnées de telle manière que les premières règles traitent ce qu'on considère comme des cas particuliers (p. ex. la cataphore) et les règles finales traitent les cas ls plus généraux.

Convention de notation dans les exemples. Dans les divers exemples qui seront présentés, on notera les couples (e_i, e_j) caractérisés par les règles en plaçant entre crochets les expressions considérées et en les marquant d'une lettre. Une lettre x en haut à droite indique que l'expression est antécédent de l'expression (ou éventuellement des expressions) qui est indicée par la même lettre en bas à droite. Une expression pronominale pouvant avoir n antécédents possibles sera marquée par une suite de n indices séparés par « / », chaque indice renvoyant à un des antécédents.

À titre d'exemple, pour la phrase suivante, les règles sur les zones d'antécédence identifieront deux antécédents possibles pour le déterminant possessif *sa* : *privatisation*, d'une part, *GAN*, d'autre part.

(8) Contrairement à $[\text{sa}]_{i/j}$ filiale bancaire, le CIC, la $[\text{privatisation}]^i$ du $[\text{GAN}]^j$ n'aura pas provoqué de désistements surprises au jour de la remise des offres fermes.

Le partage d'un indice au niveau d'une reprise et d'un antécédent possible représente une relation `coref`. Pour notre exemple, on a les deux relations :

```
coref(sa,privatisation)
coref(sa,GAN)
```

10.2.1 Pronoms clitiques

Comme nous l'avons dit, notre stratégie, dans les règles sur les zones d'antécédence, est de rendre compte des cas particuliers par un ensemble de règles qui s'appliquent en premier, et de décrire ensuite le ou les cas plus généraux comme étant ceux où aucune règle décrivant un cas particulier ne s'applique.

Cela étant, les différentes règles pour les pronoms clitiques viseront à rendre compte des cas suivants, dans l'ordre des règles :

a. le pronom apparaît dans une tournure interrogative ou apparentée (règle Z-PC.1) ;
b. le pronom est sujet d'une proposition incise (règles Z-PC.2 et Z-PC.3) ;
c. le pronom reprend un syntagme nominal en position topique (règle Z-PC.4) ;
d. l'antécédent du pronom suit le pronom — on parle alors de « cataphore » (règle Z-PC.5) ;
e. le pronom se trouve entre le sujet du verbe principal et le verbe principal (règle Z-PC.6) ;
f. le pronom dépend du verbe principal (règles Z-PC.7 et Z-PC.8) ;
g. le pronom est coréférent avec une expression qui le précède dans la même phrase ; c'est le cas général (règle Z-PC.9).

Ces formulations ne sont qu'une première approximation, les règles complètes posant parfois des conditions supplémentaires. On voit que les premières règles (Z-PC.1 à Z-PC.6) décrivent des cas très particuliers. Sur notre corpus d'étude, les règles Z-PC.1 à Z-PC.6 ne couvrent qu'un peu moins de 10 % des pronoms clitiques.

Chacune des règles sera décrite en détail dans la suite de la présente section. Avant d'en venir à cette description, nous présentons quelques conditions générales.

Conditions générales

Certaines conditions peuvent être qualifiées de générales, dans le sens où elles se retrouveront dans la majorité des règles définissant la zone d'antécédence pour les pronoms clitiques.

CONDITIONS SUR LES TRAITS ASSOCIÉS À L'EXPRESSION PRONOMINALE. L'expression pronominale (mettons `#1`) dont on cherche les antécédents possibles est un pronom clitique référentiel qui n'est pas indéfini (on exclut par là le pronom *on*). En ce qui concerne les traits qui lui sont associés, l'expression sera donc caractérisée comme suit :

```
#1[pron,clit,indef:~]
```

Par ailleurs, on exigera qu'il existe une relation `ref(#1)` ou, éventuellement, `dans(?,#1)` s'il est nécessaire de faire référence à la phrase dans laquelle figure le pronom.

SUCCESSION DE PHRASES. Les antécédents des expressions pronominales que nous cherchons à interpréter se trouveront en règle générale soit dans la même phrase que l'expression pronominale, soit dans la phrase qui précède. Pour être précis, lorsque nous ferons référence à une phrase `#1` comme « la phrase qui précède » une phrase `#2`, nous parlerons de la phrase qui précède immédiatement `#2` ou de la deuxième phrase précédant `#2` si la phrase qui précède immédiatement `#2` est une phrase nominale. Nous utiliserons donc une expression régulière de la forme suivante :

```
ST#1,(ST[noverb]),ST#2
```

Cette expression décrit une séquence de deux nœuds « phrase », entre lesquels s'intercale optionnellement un autre nœud phrase ayant le trait `noverb:+` [as-7.3.8. Dans notre terminologie, la « phrase qui précède `#2` est celle qui instancie `#1` dans cette expression.

VERBE DONT DÉPEND LE PRONOM. Pour la plupart des règles, la position du pronom dans la phrase sera exprimée en fonction de sa relation avec un verbe de la phrase. Étant donné un verbe `#2`, notre pronom `#1` sera soit sujet de `#2`, soit argument de `#2`. Nous ferons également usage de la notion de « verbe satellite » d'un autre verbe, notion définie ci-après, et nous dirons que le pronom peut être aussi sujet ou argument d'un verbe `#3` qui est « satellite » du verbe `#2`. On aura donc :

```
subj(#2,#1)
| varg(#2,#1)
| ( sat(#2,#3) & ( subj(#3,#1) | varg(#3,#1) ) )
```

La notion de « verbe satellite » est exprimée par la relation `sat`. Pour résumer

les trois options exprimées ici, nous dirons d'un pronom `#1` qui satisfait l'une des trois conditions qu'il « dépend » du verbe `#2`.

Verbe satellite. La règle définissant la relation `sat` est référencé SAT. Un verbe `#2` est dit « satellite » d'un autre verbe `#1` s'il remplit l'une des cinq conditions suivantes et s'il n'est pas coordonné à un autre verbe (conditions posée par la ligne 9) :

(a) il est un verbe à l'infinitif sans préposition argument du verbe `#1`, p. ex. *partir* dans *il parle de partir* (l. 1) ;
(b) il est complément d'un verbe dont le sujet est un pronom *il* impersonnel, p. ex. *vienne* dans *il faut qu'il vienne* (l. 2-3) ;
(c) il est complément d'un adjectif, qui est lui-même complément du verbe `#1`, ce dernier étant une copule [as-247] et ayant pour sujet un pronom impersonnel, p. ex. *prendre* dans *il est possible de le prendre* (l. 4-6) ;
(d) il est complément du verbe *être*, dont le sujet est un pronom démonstratif neutre (l. 7-8) [as-243] ;

```
1   if ( ( varg(#1,#2[verb,inf])
2          | ( varg(#1,#2[verb])
3              & subj[imperso](#1,?) )
4          | ( varg(#1[copule],#3[adj])
5              & adjarg(#3,#2[verb])
6              & subj[imperso](#1,?) )
7          | ( varg(#1[form:fetre],#2[verb])
8              & subj(#1,?[pron,dem,neutre]) ) )
9        & ~coorditems(?,?,#2)
10     )
11  sat(#1,#2)
```

Règle SAT – Verbe satellite d'un autre verbe.

L'exemple suivant illustre la condition (a). Le verbe *rassembler* est satellite du verbe *(s')efforcent*. Le pronom *les* est complément du verbe *rassembler*, mais nous dirons aussi qu'il dépend du verbe *(s')efforcent*.

(9) Le cabinet Deminor d'une part, et Me Modrikamen d'autre part, s'*efforcent* de les *rassembler*.

L'exemple suivant illustre la condition (b). Le verbe *slalome* est satellite du verbe *faut*. Le second pronom *il* est sujet de *slalome*, mais selon notre terminologie, il dépend aussi de *faut*.

(10) Mais il *faut* qu'il *slalome* autour des dossiers les plus sensibles.

Cet exemple est une paraphrase de la phrase suivante, extraite de notre corpus :

(11) Mais il lui faut aussi slalomer autour des dossiers les plus sensibles [...].

Dans cette phrase, le pronom *lui* est complément de *faut*. La notion de verbe satellite nous permet de faire abstraction des différences qui apparaissent dans la structure des constructions verbales de ces deux phrases. Dans les deux cas, nous avons un pronom qui « dépend » du verbe principal.

La condition (c) est illustrée par l'exemple suivant :

(12) Mais, de source française, on estime qu'« il *devenait impossible* de *continuer* [...] ».

Le verbe *continuer* est satellite du verbe *devenait*.

Pour finir considérons l'exemple suivant. Le verbe *renforcer* est satellite de *vise*, le verbe *déposer* est satellite de *obliger*, mais *obliger* n'est pas satellite du verbe *vise*, bien qu'il en soit complément, parce qu'il est coordonné au verbe *renforcer*.

(13) Ce décret *vise* à *renforcer* les contrôles sur les administrateurs et en particulier à les *obliger* à *déposer* leurs fonds auprès de la Caisse des dépôts.

Ainsi, si on peut dire dans les exemples (9) et (10) que le pronom dépend du verbe principal, ce n'est pas le cas pour le pronom *les* dans cette phrase. Cela aura une influence dans la manière dont nous déterminerons la zone de texte où devrait se trouver l'antécédent du pronom. Dans le cas de *les*, l'antécédent devrait se trouver dans la même phrase ; si on avait un pronom dépendant du verbe *renforcer*, son antécédent serait plutôt dans la phrase précédente.

Tournure interrogative

RÈGLE Z-PC.1. Lorsque, dans une tournure interrogative ou dans une phrase commençant par *peut-être, sans doute*, etc., un pronom clitique sujet à droite d'un verbe V est redondant par rapport à un sujet du même verbe placé à gauche, l'analyseur syntaxique identifie une relation `subj` entre le verbe et son sujet gauche et une relation `subjclit` entre le verbe et son sujet droit [as-256]. Le lien de coréférence entre les deux sujets est donc directement déductible de ces deux relations.

```
1  if ( subj(#2,#1)
2       & subjclit(#2,#3[pron,clit,indef:~])
3       & ref(#1)
4       & ref(#3)
5     )
6  coref[resolu=+](#3,#1)
```

RÈGLE Z-PC.1 – Pronom dans une tournure interrogative.

Sujet d'incise

Pour une définition des propositions incises, voir [as-7.3.7].

RÈGLE Z-PC.2. Un pronom clitique `#3` sujet à droite d'un verbe de discours [as-248] (l. 3) et qui n'est pas redondant dans une tournure interrogative (l. 4) peut être coréférent avec une expression dénotante `#4` de la phrase précédente (l. 5) qui soit sujet (l. 6) et soit (l. 7) :

- une expression ayant pour noyau un nom commun susceptible de décrire une personne [as-240] ;
- ou un nom propre véritable [as-234] ;
- ou un pronom.

Avec ces dernières conditions sur les traits associés à l'antécédent possible `#4`, on vise à identifier une expression qui puisse dénoter une personne. Les conditions sur le fait qu'on parle ici d'expressions qui se trouvent dans deux phrases successives sont exprimées par l'expression régulière et la référence à des relations `dans`.

L'assignation du trait `resolu:+` aux relations créées par cette règle permettra de tester par la suite si, pour un candidat reprise, il existe déjà une ou plusieurs relations `coref` créées par une règle précédente.

```
1  | ST#2,(ST[noverb]),ST#1 |
2  if ( dans(#1,#3[pron,clit,indef:~])
3       & subj[right](?[dicendi],#3)
4       & ~coref[resolu](#3,?)
5       & dans(#2,#4)
6       & subj(?,#4)
7       & ( #4[person] | #4[proper] | #4[pron] )
8  )
9  coref[resolu=+](#3,#4)
```

RÈGLE Z-PC.2 – Pronom clitique sujet d'incise (1).

EXEMPLE. Étant donné le texte suivant, le pronom *il* sujet de *précise* dans la second phrase peut être coréférent avec le syntagme *Le groupe* ou le pronom *il* dans la phrase précédente. La règle Z-PC.2 s'applique deux fois. Notons qu'à ce stade de l'analyse, on ne dispose pas de l'information selon laquelle *Le groupe* et *il* dans la première phrase ont la même dénotation.

(14) Le [groupe]^i Paribas va céder la participation de 25 % qu'[il]^j détient dans la banque d'affaires russe United Financial Group (UFG) au management de cette dernière. Cette décision, précise-t-[il]$_{i/j}$, lui permettra de

conduire le développement de ses activités en Russie dans le cadre de son organisation mondiale par métier.

RÈGLE Z-PC.3. Si, pour un pronom clitique #3 sujet dans une incise, la règle précédente ne s'applique pas (condition exprimée par la ligne 3), alors, l'antécédent #4 du pronom est une expression qui satisfait les mêmes conditions que précédemment à ces deux différences près :

- l'antécédent ne doit pas être un pronom ;
- l'antécédent se trouve dans la deuxième phrase avant le pronom #3.

```
1  | ST#2,ST,ST#1 |
2  if ( dans(#1,#3[pron,clit,indef:~])
3       & ~coref[resolu](#3,?)
4       & subj[right](?[dicendi],#3)
5       & dans(#2,#4)
6       & subj(?,#4)
7       & ( #4[person] | #4[proper] )
8  )
9  coref[resolu=+](#3,#4)
```

RÈGLE Z-PC.3 – Pronom clitique sujet d'incise (2).

EXEMPLE. La règle Z-PC.3 s'applique deux fois pour le texte suivant. Le pronom *il* dans la dernière phrase peut avoir pour antécédent — abstraction faite des contraintes d'accord qui s'appliqueront par la suite — soit *les parlementaires*, soit *Arnaud Montebourg*.

(15) Lorsque les [parlementaires]^i ont abordé la délicate question du barème facturé par les mandataires, [Arnaud Montebourg]^j a suscité des remous dans la salle. « Dans certains ressorts, cette facturation donne lieu à une bataille permanente avec les mandataires. Ici, ce n'est pas le cas car on est entre amis », a-t-[il]$_{i/j}$ souligné avec un brin de provocation.

Notons que la seconde phrase du texte constitue ici la première partie du discours rapporté qui se poursuit dans la troisième phrase et se termine par la proposition incise dont *il* est le sujet. Une meilleure formulation de ces deux règles pourrait être de dire que le pronom reprend un antécédent situé dans la phrase (ou la proposition dans certains cas) qui précède immédiatement l'unité que constitue l'ensemble du discours rapporté. L'analyseur syntaxique ne donne pas cette information sur le discours rapporté et nous nous en tenons donc à la formulation proposée. Par ailleurs, l'exemple (14) montre qu'il n'y a pas toujours de discours rapporté explicite.

Source détachée en début de phrase

Règle Z-PC.4. Un autre cas particulier, traité comme tel par une règle spécifique, est celui où un pronom clitique a pour antécédent un syntagme nominal détaché en début de phrase, comme dans :

(16) Ta sœur, elle est merveilleuse.

Grevisse [37, §448] note à ce propos que « dans le cas des sujets, des compléments essentiels ou des attributs, [les] détachements entraînent ordinairement la redondance, c'est-à-dire la présence d'un pronom devant le verbe. »

La règle décrite ici vise à rendre compte des cas où un pronom renvoie à un antécédent détaché en début de phrase uniquement. On fait l'hypothèse que le détachement en fin de phrase appartient plutôt au langage parlé.

Par ailleurs, les locutions prépositionnelles *quant à* et *pour ce qui est de* sont également utilisées pour mettre en relief une expression, d'une manière similaire au détachement à gauche du syntagme nominal (voir [37, §1044d]). La règle Z-PC.4 rend également compte de ces cas.

```
1  | ST#9{?*[punct],
2         SC{?*[punct],
3            NP{?*,#2[last]};PP{?[form:fquanta],
4                               NP{?*,#2[last]}},
5            ?*[fonc:~fsubj],
6            PUNCT[form:fcm],
7            NP[fonc:fsubj]{?*,#3[last]},
8            ?*,
9            FV{?*,#1[last]}}} |
10 if ( subj(#1,#3)
11      & dans(#9,#4[pron,clit,indef:~])
12      & ( (#4 :: #3)
13          | varg(#1,#4)
14          | ( sat(#1,#5) &
15              ( subj(#5,#4) | varg(#5,#4) ) ) )
16      & ref(#2)
17      & ( #2[proper] | determ(?[deter],#2) )
18      & ~subj(?,#2)
19    )
20 coref[resolu=+](#4,#2)
```

Règle Z-PC.4 – Pronom clitique avec source détachée en début de phrase.

Deux exemples extraits de notre corpus d'étude pour lesquels la règle Z-PC.4 s'applique.

(17) Toutes ces [formes]^i de propriété admises au-dessus de la propriété civile, nous aurions simplement pu [les]$_i$ nommer « propriété privée » [...]

(18) Quant à la [neutralisation]^i, [elle]$_i$ serait limitée à la période janvier à octobre 1999 et ne représenterait qu'un peu plus d'un milliard.

La règle comprend une expression régulière relativement complexe. On a une phrase `#9` dont la proposition principale débute [3] par

- un syntagme nominal dont le noyau est `#2`
- ou un syntagme prépositionnel commençant par la locution *quant à* ou *pour ce qui est de* [as-247] et dont le noyau est `#2`

On exige (l. 17) de l'expression `#2` qu'elle soit un nom propre ou qu'elle soit déterminée par un article défini, démonstratif ou possessif (catégorie définie par l'attribut général `deter` [as-244]). Par ailleurs, cette expression ne doit pas être sujet (l. 18).

L'expression `#2` est suivie d'une suite de 0 à n nœuds qui n'ont pas fonction de sujet, puis d'une virgule, puis d'un syntagme nominal dont le noyau est `#3`. Cette dernière expression est sujet du verbe `#1`, verbe noyau de la proposition principale (l. 10).

Dans ce contexte, si on a un pronom `#4` qui dépend du verbe principal (l. 12-15), alors ce pronom peut être coréférent avec l'expression `#2`, l'expression détachée à gauche.

Notons que, compte tenu des conditions qui seront posées par la suite sur la non-existence de relations ayant ce trait, l'assignation du trait `resolu:+` à la relation `coref` créée revient à dire que le pronom peut être coréférent *seulement* avec cette expression.

Cataphore

La règle Z-PC.5 vise à rendre compte des cas de « cataphore », c'est-à-dire des cas particuliers où un pronom clitique renvoie à un antécédent qui le suit plutôt qu'à un antécédent qui le précède.

Le contexte dans lequel peuvent se produire les cas de cataphore est décrit comme celui où une proposition finie noyau contient une proposition finie noyau enchâssée, qui ne soit pas une incise (`SC[disco:~]`) [as-7.3.7] et qui ne contienne pas de coordination. La proposition enchâssée ne doit pas être précédée d'une autre proposition, ni d'un syntagme ayant fonction de sujet (l. 1) [as-253]. On identifie par `#1` le verbe de la proposition enchâssée (l. 2). Le verbe de la proposition qui contient l'enchâssée est identifié par `#2` (l. 7). Son sujet `#5` (l. 5 et 14) est précédé d'une virgule (l. 4) [as-253].

[3] Abstraction faite d'éventuels symboles de ponctuation ou balises SGML, qui instancieraient `?*[punct]`.

Dans ce contexte, un pronom #3 qui dépend du verbe #1 (l. 8-11), peut être coréférent avec une expression dénotante #6 qui précède le verbe #2 (l. 16) et qui est soit le sujet de #2 soit une expression qui suit le sujet de #2 (l. 17). Notons que si elle n'est pas le sujet, alors cette expression est *a priori* soit un pronom clitique argument du verbe, soit un complément du sujet.

Cette règle ne s'applique pas si une des règles précédentes s'est appliquée (l. 12).

```
1   | SC{?*[sc:~,fonc:~fsubj],
2        SC[disco:~]{?*[coord:~],FV[last]{?*,#1[last]}},
3        ?*,
4        PUNCT[form:fcm],
5        NP[fonc:fsubj]{?*,#5[last]},
6        ?*,
7        FV[last]{?*,#2[last]}} |
8   if ( ( subj(#1,#3)
9          | varg(#1,#3)
10         | ( sat(#1,#4) &
11             ( subj(#4,#3) | varg(#4,#3) ) ) )
12       & dans(#8,#3[pron,clit,indef:~])
13       & ~coref[resolu](#3,?)
14       & subj(#2,#5)
15       & dans(#8,#6)
16       & (#6 < #2)
17       & ( (#5 : #6) | (#5 < #6) )
18     )
19  coref[resolu=+](#3,#6)
```

RÈGLE Z-PC.5 – Pronom clitique. Contexte de cataphore.

EXEMPLES. Dans les deux exemples suivants, nous avons délimité les propositions noyau par « SC{ » à gauche et « } » à droite. Dans la phrase suivante, le pronom clitique *lui* est dans une proposition enchâssée dans la proposition principale noyau de la phrase (qui va du début de la phrase au verbe *garde*). Il peut être coréférent avec *Bercy*.

(19) SC{ Pris en tenaille entre les contraintes de Bruxelles, SC{ qui dans un certain sens [lui]$_i$ permettent } de justifier les réformes vis-à-vis des établissements concernés, et la pression des banques AFB SC{ qui jugent } le mouvement trop lent, [Bercy]^i se garde } bien de clamer SC{ qu'il a } en tête un plan d'ensemble.

La règle Z-PC.5 s'applique deux fois pour la phrase suivante. Le pronom clitique *lui* peut être coréférent soit avec *la réputation de la banque*, soit avec *la banque*.

(20) `SC{` Il est `}` vrai `SC{` que `SC{` si certains [lui]$_{i/j}$ reconnaissent `}` un réseau intéressant dans sa région et une clientèle fidèle, la [réputation]^{i} de la [banque]^{j} est `}` pour le moins écornée par des années d'errements.

Le pronom est entre le sujet et le verbe principal

La règle Z-PC.6 rend compte du cas très particulier où un pronom clitique se trouve entre le sujet du verbe principal et le verbe principal sans pour autant être dans une proposition noyau. Il s'agit donc *a priori* d'un pronom clitique argument d'un verbe à l'infinitif. Dans ce cas, l'antécédent du pronom est une expression qui précède le pronom soit dans la même phrase que celui-ci, soit dans la phrase précédente.

Les contraintes posées par l'expression régulière sont les suivantes. On a une phrase `#2` dont la proposition principale a pour noyau le verbe `#3` (l. 8). Ce verbe a pour sujet `#5` (l. 4 et 9). Dans la proposition, il ne doit pas y avoir avant ledit sujet de proposition ni de syntagme ayant fonction de sujet (l. 3). Entre le sujet et le verbe se trouve un nœud qui domine un pronom `#4`, ce nœud ne doit pas être une proposition ni une insertion (l. 6). Enfin entre le sujet et le nœud qui domine le pronom, il ne doit pas y avoir de coordination, de virgule, d'insertion ou de proposition (l. 5).

On le voit, le contexte d'application de cette règle est très restreint. Dans le cas où un pronom se trouverait entre le sujet et le verbe de la principale, sans que la phrase satisfasse les contraintes posées par cette règle, la règle qui s'appliquerait serait la règle Z-PC.9 (cas général), c'est-à-dire que l'antécédent du pronom serait dans la même phrase que le pronom.

Exemples. L'exemple suivant illustre le cas où le pronom renvoie à une expression de la phrase précédente. L'application de la règle donne lieu à l'identification de quatre antécédents possibles [4].

(21) Sans passage au [civilisme]^{i}, il est impossible de surmonter les nouvelles [tentatives]^{j} de réalisation de l'[idéologie]^{k} communiste. En outre, la [nécessité]^{l} de [les]$_{i/j/k/l}$ surmonter est évidente.

L'exemple suivant illustre le cas où le pronom renvoie à un antécédent figurant dans la même phrase. Six antécédents possibles sont identifiés pour le pronom clitique *la*.

[4] Une nouvelle fois, nous faisons abstraction des contraintes qui seront décrites par la suite. Cet exemple est une adaptation d'un extrait du corpus.

```
1  | ST#1,(ST[noverb]),
2    ST#2{?*[punct],
3         SC{?*[sc:~,fonc:~fsubj],
4            NP[fonc:fsubj]{?*,#5[last]},
5            ?*[coord:~,form:~fcm,ins:~,sc:~],
6            ?[ins:~,sc:~]{?*,PRON^#4[clit,indef:~]},
7            ?*,
8            FV{?*,#3[last]}}} |
9  if ( subj(#3,#5)
10       & ref(#4)
11       & ( dans(#1,#6) | dans(#2,#6) )
12       & ( #6 < #4 )
13     )
14 coref[resolu=+](#4,#6)
```

RÈGLE Z-PC.6 – Pronom clitique entre le sujet et le verbe principal

(22) La [différence]^i est de taille, [l'on]^j peut dire qu'[elle]^k a fait époque. La [formation]^l post-socialiste avec l'[étatisation]^m de la [propriété]^n socialiste pour [la]$_{i/j/k/l/m/n}$ transformer en propriété privée est, du point de vue des coordonnées socio-historiques, clairement dans une situation pré-capitaliste [...]

Le pronom dépend du verbe principal

Un cas plus fréquent que les cas particuliers que nous avons vus, mais néanmoins moins fréquent que le cas général que nous décrirons après, est celui où un pronom clitique dépend du verbe principal d'une phrase. Dans ce cas, on a deux possibilités, dont on rend compte respectivement par les règles Z-PC.7 et Z-PC.8 :

- soit la proposition principale noyau contient une enchâssée, auquel cas le pronom renvoie à une expression qui dépend du verbe de cette enchâssée ;
- soit la proposition principale noyau ne contient pas d'enchâssée, auquel cas le pronom renvoie à une expression de la phrase précédente.

Ces deux règles ne présentent pas de difficulté notable. Dans les deux cas, on a un pronom clitique référentiel `#4` qui dépend d'un verbe `#1`, verbe principal de la phrase ; aucune règle ne doit s'être appliquée précédemment pour ce pronom (l. 5-10 pour Z-PC.7, l. 3-8 pour Z-PC.8).

Dans le cas où la principale contient une enchâssée (Z-PC.7), on exige en outre que le pronom n'ait pas lui-même été identifié comme l'antécédent possible

```
| ST{?*[punct],SC{?*,
                    SC{?*,FV{?*,#2[last]}},
                    ?*,
                    FV{?*,#1[last]}}} |
if ( ( subj(#1,#4)
        | varg(#1,#4)
        | ( sat(#1,#5) &
            ( subj(#5,#4) | varg(#5,#4) ) ) )
      & ref(#4[pron,clit,indef:~])
      & ~coref[resolu](#4,?)
      & ~coref(?,#4)
      & connect(#2,#7[conj])
      & ( subj(#2,#8)
          | varg(#2,#8)
          | ( sat(#2,#6) &
              ( subj(#6,#8) | varg(#6,#8) ) ) )
      & ref(#8)
    )
coref[resolu=+](#4,#8)
```

Règle Z-PC.7 – Pronom clitique dépendant du verbe principal (1)

```
| ST#3,(ST[noverb]),ST#2{?*[punct],
                           SC{?*,FV{?*,#1[last]}}} |
if ( ( subj(#1,#4)
        | varg(#1,#4)
        | ( sat(#1,#5) &
            ( subj(#5,#4) | varg(#5,#4) ) ) )
      & ref(#4[pron,clit,indef:~])
      & ~coref[resolu](#4,?)
      & dans(#3,#6)
    )
coref(#4,#6)
```

Règle Z-PC.8 – Pronom clitique dépendant du verbe principal (2)

```
if ( dans(#1,#2[pron,clit,indef:~])
      & ~coref[resolu](#2,?)
      & dans(#1,#3)
      & (#3 < #2)
    )
coref(#2,#3)
```

Règle Z-PC.9 – Pronoms clitiques. Cas général.

d'une expression (l. 11). On vise par là à éviter une circularité dans les cas de cataphore. Pour la phrase suivante,

(23) Bien qu'*ils* expriment des critiques souvent concrètes et spécifiques, *ils* jouent également de prudence.

la règle Z-PC.5 identifiera une relation `coref` entre le premier *ils* et le second. Il convient d'éviter maintenant d'interpréter le second *ils* par rapport au premier.

On exige de la proposition enchâssée dans la principale qu'elle soit introduite par une conjonction (l. 12). Les propositions relatives sont donc exclues.

Enfin l'antécédent du pronom est une expression dénotante `#8` qui dépend du verbe noyau de l'enchâssée (l. 13-17).

Pour la règle Z-PC.8, la condition sur la source est simplement qu'elle soit une expression dénotante de la phrase précédente. Il est à noter que dans cette règle le trait `resolu:+` n'est pas assigné aux relations `coref` créées, si bien que, par la règle suivante, il n'est pas exclu que le pronom ait son antécédent dans la même phrase.

Exemples. L'exemple suivant illustre une application de la règle Z-PC.7. Le pronom *elle* a pour antécédent *l'opération* dans la proposition enchâssée dans la principale.

(24) `SC{` `SC{` Si l'[opération]^i aboutit `}`, [elle]$_i$ sera `}` la plus importante prise de contrôle par un groupe étranger d'une banque brésilienne.

L'exemple suivant illustre une application de la règle Z-PC.8. Le pronom *la* est argument du verbe *laisser*, qui est satellite du verbe principal *devrait* ; autrement dit le pronom « dépend » de *devrait*. Son antécédent est dans la phrase précédente.

(25) [...] la [banque]^i se voit forcément limitée dans [sa]^j [volonté]^k d'augmenter encore [ses]^l [parts]^m de marché. La Réserve fédérale ne devrait pas en effet [la]$_{i/j/k/l/m}$ laisser franchir la barre des 10 % des dépôts bancaires.

Cas général

La règle Z-PC.9 rend compte du cas général pour l'interprétation des pronoms clitiques. Si aucune des règles précédemment définies ne s'applique, un pronom clitique est coréférent avec une expression qui le précède dans la même phrase.

10.2.2 Pronoms disjoints

Les règles sur les zones d'antécédence pour les pronoms disjoints sont au nombre de trois. Les deux premières traitent les cas où l'antécédent du pronom se trouve dans la phrase précédente. Dans un cas le pronom modifie le sujet d'une

proposition incise ou d'une proposition principale (Z-PD.1), dans l'autre, il est complément du verbe principal ou précède le sujet du verbe principal (Z-PD.2). La troisième règle rend compte des cas où l'antécédent du pronom se situe dans la même phrase que le pronom. Le pronom suit alors le sujet du premier verbe (Z-PD.3).

Pronom complément du sujet d'une proposition incise ou principale

Lorsqu'un pronom disjoint est complément du sujet d'une proposition principale ou d'une proposition incise, son antécédent se situe dans la phrase précédente, ou dans l'une des deux phrases précédentes si la phrase précédente est une phrase nominale (règle Z-PD.1).

Les conditions sur le pronom sont les suivantes. On a une phrase `#2` dont le verbe principal est `#3` (l. 1-2). Le pronom est une expression `#4` dans cette phrase (l. 3). On a par ailleurs une expression dénotante `#5` également dans la phrase `#2` (l. 4). Cette expression est soit le sujet du verbe principal, soit sujet d'une proposition incise (l. 5). Le pronom est complément du sujet `#4` dans l'une ou l'autre de ces deux configurations (l. 6-7) :

- le pronom est introduit par la préposition *d'entre* (cf. l'exemple (26) ci-dessous) [5] ;
- ou le sujet (`#4`) est un pronom et le pronom disjoint est introduit par la préposition *de* (cf. l'exemple (27) ci-dessous).

Exemples. Les deux exemples suivants illustrent chacun l'une des deux configurations possibles en ce qui concerne la relation du pronom à l'expression sujet. Dans l'exemple (26) le pronom est complément du sujet de la principale ; dans l'exemple (27), il complément du sujet d'une incise.

(26) En réalité, lors de la dernière [réunion]^i de la [CNDA]^j le [23 mars]^k dernier, six [administrateurs]^l avaient persisté dans [leur]^m [refus]^n de payer, alors que la [Commission]^o [les]^p menaçait de sanctions disciplinaires. L'un d'entre [eux]$_{i/j/k/l/m/n/o/p}$ avait immédiatement fait appel [...].

(27) Interrogés sur [leurs]^i [rémunérations]^j, les [mandataires]^k liquidateurs du [Var]^l affirment facturer une [moyenne]^m de 20 000 francs par dossier, ce qui ne [leur]^n semble pas excessif, compte tenu des [démarches]^o nombreuses effectuées. « Souvent, nous nous transformons en assistantes sociales pour les débiteurs, affirme l'un d'[eux]$_{i/j/k/l/m/n/o}$, et il faut répéter les explications auprès de créanciers qui n'y comprennent rien. »

[5] La suite *d'entre* est considérée comme une seule unité lexicale par l'analyseur [as-247].

```
| ST#1,(ST[noverb]),ST#2{?*[punct],
                          SC{?*,FV[last]{?*,#3[last]}}} |
if ( dans(#2,#4[pron,ton])
     & dans(#2,#5)
     & ( subj(#3,#5) | subj[right](?,#5) )
     & ( nmod(#5,?[form:fdentre],#4)
         | nmod(#5[pron],?[form:fde],#4)
       )
     & dans(#1,#6)
   )
coref[resolu=+](#4,#6)
```

Règle Z-PD.1 – Pronom disjoint complément du sujet d'une incise ou principale

```
| ST#1,(ST[noverb]),ST#2{?*[punct],
                          SC{?*,FV[last]{?*,#3[last]}}} |
if ( dans(#2,#4[pron,ton])
     & subj(#3,#5)
     & ( (#4 < #5)
         | varg(#3,?,#4)
         | vmod(#3,?,#4) )
    & dans(#1,#6)
   )
coref(#4,#6)
```

Règle Z-PD.2 – Pronom disjoint complément du verbe principal ou précédant le sujet

```
|! ST#2{?*,FV^{?*,#3[last]}} |
if ( dans(#2,#4[pron,ton])
     & ~coref[resolu](#4,?)
     & subj(#3,#5)
     & (#5 < #4)
     & dans(#2,#6)
     & (#6 < #4)
     & ( (#5 < #6) | (#5 :: #6) )
   )
coref(#4,#6)
```

Règle Z-PD.3 – Pronom disjoint après le sujet du premier verbe

Pronom disjoint complément du verbe principal ou précédant le sujet

La règle Z-PD.2 rend compte d'une deuxième situation dans laquelle un pronom disjoint peut avoir son antécédent dans la phrase précédente.

On a une phrase `#2`, dont le verbe principal est `#3`. Le sujet du verbe principal est `#5`. Si un pronom disjoint `#4`

- précède le sujet `#5` (l. 5) ;
- ou est complément du verbe `#3` et est introduit par une préposition (l. 6-7),

alors ce pronom peut être coréférent avec une expression de la phrase précédente ou de l'une des deux phrases précédentes si la phrase précédente est une phrase nominale.

Il est à noter que cette règle, contrairement à la précédente, n'exclut pas une application de la règle suivante.

Exemples. Les conditions posées par les lignes 4, 5 et 6 sont illustrées respectivement par les trois exemples suivants.

(28) Le [caractère]^i anti-juridique, anti-libéral de [leur]^j [utopie]^k est lié à cela. Pour [eux]$_{i/j/k}$ pratiquement, la propriété égale est l'idéal cherché [...] [6].

(29) National Mutual et Lend Lease ne fusionneront pas : [Axa]^i doit faire une [croix]^j sur [son]^k [projet]^l. Cette opération aurait fait de [lui]$_{i/j/k/l}$ l'actionnaire de contrôle d'un des premiers groupes financiers australiens.

(30) A titre d'exemple, le [consommateur-citoyen]^i décide d'acheter en fonction de l'[utilité]^j sur la [base]^k d'une [estimation]^l personnelle coût-bénéfice, mais aussi en fonction d'une [connaissance]^m des [répercussions]^n de [son]^o [acte]^p économique sur les [circuits]^q de production-consommation aux [incidences]^r sociales, culturelles et écologiques, générales. Cela implique pour [lui]$_{i/j/k/l/m/n/o/p/q/r}$ une prise de connaissance diversifiée, [...].

Pronom disjoint après le sujet du premier verbe d'une phrase

La règle Z-PD.3 rend compte des cas où l'antécédent d'un pronom disjoint se trouve dans la même phrase que le pronom.

Étant donné

- une phrase `#2` dont le premier verbe est `#3` — ce qu'exprime l'expression régulière restreinte à la plus courte instanciation (symbole `!`, voir p. 273) ;
- un pronom disjoint `#4` dans la phrase `#2` pour lequel la règle Z-PD.1 ne s'est pas appliquée (l. 2-3) ;
- `#5` le sujet de `#3`, qui précède le pronom `#4` (l. 4-5)

[6] Dans ce texte, le pronom *eux* est coréférent avec le déterminant possessif *leur*. Si, *a priori*, le fait d'autoriser qu'un pronom renvoie à un déterminant possessif peut paraître discutable, cet exemple justifie cette pratique.

le pronom #4 peut être coréférent avec une expression #6 figurant dans la phrase #2 (l. 6), expression telle qu'elle précède le pronom (l. 7) et est soit précédée de #5, soit identique à #5.

EXEMPLE. Dans la phrase suivante, l'expression *administrateurs* est le seul antécédent possible pour le pronom *eux*.

(31) Reste quatre [administrateurs]^i : parmi [eux]$_i$, trois ont démissionné de leur fonction en début d'année.

10.2.3 Déterminants possessifs

Les règles sur les zones d'antécédence des déterminants possessifs sont au nombre de quatre.

Les deux premières règles traitent deux cas où l'antécédent peut se trouver dans la phrase précédente, d'une part lorsque le déterminant se trouve entre le sujet du verbe principal et le verbe principal — ce contexte étant restreint par ailleurs d'une manière qui sera décrite ci-après, d'autre part lorsque le déterminant possessif détermine le sujet de la proposition principale.

La troisième règle décrit les cas de cataphore.

Enfin, la dernière règle rend compte du cas général pour les déterminants possessifs, qui est que le déterminant renvoie à une expression qui le précède dans la même phrase.

Déterminant possessif entre le sujet et le verbe principal

La règle Z-DP.1 traite un cas très particulier. On a une phrase #2 dont le verbe principal est #3 ; le sujet du verbe principal est #5 (l. 3 et 10). Entre les expressions #5 et #3, on a un déterminant possessif #4 (l. 5). Sur cette structure, on pose quelques contraintes :

- il ne doit y avoir avant l'expression #5 ni proposition enchâssée ni syntagme ayant fonction de sujet (l. 3) ;
- il ne doit pas y avoir entre l'expression #5 et le déterminant possessif de conjonction de coordination, de virgule, d'insertion entre parenthèses, ni de proposition enchâssée (l. 5) ;
- le déterminant possessif ne doit pas être dans une enchâssée ou une insertion entre parenthèses (l. 6).

Dans une configuration où l'une de ces contraintes ne serait pas respectée, le déterminant possessif aurait son antécédent dans la même phrase. Dans la configuration décrite par la règle, son antécédent peut être une expression de la phrase précédente (l. 11). L'application de cette règle n'exclut pas l'application ultérieure de la règle Z-DP.4, présentée ci-après. Autrement dit, le déterminant possessif peut aussi avoir son antécédent dans la même phrase.

```
1   | ST#1,(ST[noverb]),
2     ST#2{?*[punct],
3          SC{?*[sc:~,fonc:~fsubj],
4             NP[fonc:fsubj]{?*,#5[last]},
5             ?*[coord:~,form:~fcm,ins:~,sc:~],
6             ?[ins:~,sc:~]{?*,DET^#4[poss]},
7             ?*,
8             FV{?*,#3[last]}}} |
9   if ( ref(#4)
10       & subj(#3,#5)
11       & dans(#1,#7)
12     )
13  coref(#4,#7)
```

Règle Z-DP.1 – Déterminant possessif entre le sujet et le verbe principal.

```
1   | ST#1,(ST[noverb]),ST#2{?*[punct],
2                            SC{?*[sc:~,fonc:~fsubj],
3                               NP[fonc:fsubj]{?*,#5[last]},
4                               ?*,
5                               FV{?*,#3[last]}}} |
6   if ( dans(#2,#4[det,poss])
7        & subj(#3,#5)
8        & determ(#4,#5)
9        & dans(#1,#7)
10     )
11  coref[resolu=+](#4,#7)
```

Règle Z-DP.2 – Possessif déterminant le sujet du verbe principal.

```
1  | SC{?*[sc:~,fonc:~fsubj],
2       DET^#1[poss],
3       ?*[fonc:~fsubj],
4       PUNCT[form:fcm],
5       NP[fonc:fsubj]{?*,#2[last]},
6       ?*,
7       FV{?*,#8[last]}} |
8  if ( ref(#1)
9       & subj(#8,#2)
10      & ref(#5)
11      & ( (#5 :: #2)
12          | narg(#2,?,#5)
13          | nmod(#2,?,#5)
14          | varg(#8,#5[pron]) )
15     )
16 coref(#1,#5)
```

Règle Z-DP.3 – Déterminants possessifs. Contexte de cataphore.

```
1  if ( dans(#1,#2[det,poss])
2       & ~coref[resolu](#2,?)
3       & dans(#1,#3)
4       & (#3 < #2)
5     )
6  coref(#2,#3)
```

Règle Z-DP.4 – Déterminants possessifs. Cas général.

EXEMPLES. La règle Z-DP.1 s'applique pour les deux déterminants possessifs de la phrase suivante, identifiant comme antécédents possibles pour les deux déterminants les expressions *TVA, coût, entreprises* et *solutions*, auxquelles s'ajoutent *son* et *impact* pour le déterminant *leurs*.

(32) La [TVA]^i représente un [coût]^j pour les [entreprises]^k. Dès lors, toutes les solutions pour réduire [son]$^l_{i/j/k}$ [impact]^m sur [leurs]$_{i/j/k/l/m}$ comptes sont les bienvenues.

L'exemple suivant illustre le cas où un déterminant possessif (le déterminant *sa*) renvoie en fait à une expression de la même phrase. Cet antécédent sera identifié par la règle Z-DP.4.

(33) [...] ces [personnes]^i sont propriétaires de titres pour [leur]^j [valeur]^k nominale et n'ont aucun droit sur les [réserves]^l [...]. Dans ces conditions, le prix du transfert par Cera de [sa]$_{i/j/k/l}$ participation dans le MRBB (15,1 %) est symbolique [...].

Possessif déterminant le sujet du verbe principal

Le second cas où l'antécédent d'un déterminant possessif peut se trouver dans la phrase précédente est celui où le possessif détermine le sujet de la proposition principale d'une phrase.

Une contrainte additionnelle est exprimée dans la règle Z-DP.2 :

– le sujet (`#5`) ne doit pas être précédé d'une enchâssée ou d'un syntagme ayant fonction de sujet (l. 2).

Selon notre règle, le déterminant non seulement peut, mais *doit* avoir pour antécédent une expression de la phrase précédente, ce qu'on indique par l'assignation du trait `coref:+` à la (aux) relation(s) créée(s).

EXEMPLE. Dans la seconde phrase du texte suivant, le déterminant possessif *ses* détermine le sujet de la principale. Son antécédent est dans la phrase précédente.

(34) En revanche, le [groupe]^i néerlandais ABN Amro, qui a [lui aussi]^j été approché, a choisi la [voie]^k de la [neutralité]^l. Au lendemain même du choix du Crédit Mutuel pour reprendre le CIC, [ses]$_{i/j/k/l}$ dirigeants n'ont d'ailleurs pas hésité à prendre contact avec la direction du groupe mutualiste pour évoquer, pourquoi pas, de possibles pistes de partenariat.

Contexte de cataphore

La règle Z-DP.3 décrit les cas de cataphore, c'est-à-dire les cas où un déterminant possessif peut renvoyer à une expression qui le suit, plutôt qu'à une expression qui le précède.

L'expression régulière dans cette règle décrit une proposition noyau `SC`. Le verbe noyau de la proposition est `#8` (l. 7). Le sujet de `#8` est un syntagme nominal dont le noyau est `#2` (l. 5 et 9). Ce syntagme est précédé d'une virgule (l. 4). Entre le début de la proposition et ladite virgule se trouve un déterminant possessif `#1` (l. 2). Ce déterminant peut être dominé par un nœud de n'importe quel type, mais la séquence de nœud le précédant ne doit pas contenir de nœud de type `SC` ou de nœud ayant fonction de sujet (l. 1). Par ailleurs, il ne doit pas y avoir de nœud ayant fonction de sujet entre le nœud qui domine le déterminant possessif et le sujet `#2` — on cherche à s'assurer que `#2` est le premier sujet de `#8`, c'est-à-dire un sujet non coordonné.

Dans ce contexte, le déterminant possessif `#1` peut être coréférent avec une expression dénotante `#5` telle que :

- `#5` est le sujet de `#8` (plus précisément il est identique à `#2`, l. 11) ;
- ou `#5` est complément de `#2`, sujet de `#8` (l. 12-13) ;
- ou `#5` est un pronom argument du verbe `#8`.

Exemple. Dans la phrase suivante, le déterminant possessif *sa* peut être coréférent soit avec le sujet de la principale, soit avec le complément dudit sujet.

(35) Contrairement à [sa]$_{i/j}$ filiale bancaire, le CIC, la [privatisation]^{i} du [GAN]^{j} n'aura pas provoqué de désistements surprises au jour de la remise des offres fermes.

Cas général

Le cas général, en ce qui concerne la zone d'antécédence pour un déterminant possessif, est que l'antécédent du possessif soit une expression qui le précède dans la même phrase. C'est ce qu'exprime la règle Z-DP.4. Elle s'applique pour tout déterminant possessif, à l'exception de ceux pour lesquels la règle Z-DP.2 s'applique (l. 2).

Exemples. Deux exemples. Le premier illustre un cas où le déterminant est entre le verbe principal et le sujet, mais dans un contexte tel que la règle Z-DP.1 ne s'applique pas. Le second illustre le cas d'une phrase sans proposition principale.

(36) [Celui-ci]^{i}, après avoir vu le [Credito Italiano]^{j} torpiller [son]$_{i/j}$ projet de le fusionner avec Banca di Roma et Comit [...], milite aujourd'hui pour l'alliance entre Comit et Banca di Roma.

(37) Autre [candidat]^{i} à devoir encore faire [ses]$_{i}$ preuves : Eureko.

10.3 Contraintes

Les contraintes sont des règles qui expriment l'impossibilité d'un lien de coréférence pour un couple (e_i, e_j). Dans le processus de résolution des pronoms tel que nous l'avons implanté dans XIP, les contraintes sont des règles qui visent à effacer des relations `coref` de l'univers de dénotation construit par application des règles sur les expressions dénotantes et les zones d'antécédence. Les règles — à une exception près — aboutissent donc toujours à la conclusion « ~ », qui signifie qu'une relation doit être effacée. Les conditions posées dans chaque règle contiennent donc toujours une référence à une relation `coref(#1,#2)`, préfixée par ^ pour indiquer que la relation doit être effacée.

Dans la mesure où l'effacement d'une relation `coref(#1,#2)` par application d'une contrainte n'est jamais conditionnée à l'existence d'une autre relation `coref(#1,#3)`, contrairement à ce qui se passe avec les préférences qui seront décrites plus loin, on peut voir les contraintes non comme des conditions qui seraient valables *après* application des règles sur les expressions dénotantes et les zones d'antécédence, mais comme des conditions qui doivent être respectées *en même temps* que doivent l'être les conditions posées par les règles sur les expressions dénotantes et les zones d'antécédence. L'ordre des contraintes les unes par rapport aux autres n'est pertinent ni théoriquement, ni dans l'implantation. Le fait que les contraintes s'appliquent après les règles sur les zones d'antécédences est pertinent informatiquement (cela détermine l'interprétation des formules), mais ne l'est pas théoriquement : l'ensemble des règles définies dans le présent chapitre peut être vu comme une conjonction, modulo l'ordonnancement interne des règles sur les zones d'antécédence.

Nous distinguons trois types de contraintes :

- les contraintes d'accord en genre et en nombre (C-A) ;
- les contraintes relationnelles (C-R) ;
- les contraintes sur les insertions (C-I).

Ces différentes contraintes sont présentées successivement dans les trois sections suivantes.

10.3.1 Contraintes d'accord

Accord en genre

Les contraintes d'accord en genre sont pertinentes pour les pronoms seulement, le genre des déterminants possessifs ne donnant aucun indication sur le genre de leur antécédent. Ces contraintes sont exprimées par deux règles, une pour le féminin (C-A.1), une pour le masculin (C-A.2).

Intuitivement, les règles C-A.1 et C-A.2 disent que s'il existe une relation `coref(#1,#2)` telle que le nœud qui instancie `#1` et le nœud qui instancie `#2` ne s'accordent pas en genre, alors cette relation doit être effacée. Dit autrement : l'hypothèse d'une coréférence entre ces deux nœuds doit être rejetée.

Les règles ne s'appliquent que pour des pronoms qui ne sont pas ambigus en ce qui concerne le genre : un pronom féminin non ambigu en genre est identifié par le fait qu'il a le trait `fem:+` et n'a pas le trait `masc:+`, et inversement pour un pronom masculin non ambigu en genre [7]. Aucune des deux règles d'accord en genre ne s'applique donc pour le pronom clitique *lui*, qui a à la fois le trait `fem:+` et le trait `masc:+`.

L'exigence de non-ambiguïté en genre est également valable pour l'antécédent potentiel du pronom. Deux contraintes additionnelles sont posées sur l'antécédent potentiel : les deux règles ne s'appliquent pas si l'antécédent potentiel est un déterminant possessif (exigence `det:~`).

```
1  if ( ^coref(#1,#2)
2       & #1[pron,fem,masc:~]
3       & #2[det:~,fem:~,masc]
4     ) ~
```

Règle C-A.1 – Accord en genre. Pronom féminin.

```
1  if ( ^coref(#1,#2)
2       & #1[pron,masc,fem:~]
3       & #2[det:~,masc:~,fem]
4     ) ~
```

Règle C-A.2 – Accord en genre. Pronom masculin.

Accord en nombre

Les contraintes d'accord en nombre sont exprimées par deux règles, une pour le singulier (C-A.3), une pour le pluriel (C-A.4).

Nous distinguons le nombre « grammatical », dont rendent compte les traits `sg:+` et `pl:+`, du nombre « sémantique » des expressions, dont rendent compte les traits `semsg:+` et `sempl:+` [as-245].

Les deux règles d'accord en nombre ont la même structure. On ne décrit que la première, qui concerne les reprises par une expression pronominale ayant un nombre sémantique singulier. Étant donné une relation `coref(#1,#2)` entre une expression pronominale `#1` de nombre sémantique singulier et une expression `#2`, si `#2` n'est ni grammaticalement ni sémantiquement singulier (`sg:~,semsg:~`) et

[7] Les traits rendant compte du genre sont décrit page 246.

n'est pas un nom propre, ou si `#2` est un déterminant possessif ayant un nombre sémantique pluriel, alors l'existence de la relation `coref(#1,#2)` est impossible (le système efface la relation de son univers de dénotation).

```
1  if ( ^coref(#1[semsg],#2)
2       & ( #2[sg:~,semsg:~,proper:~]
3           | #2[det,sempl] )
4     ) ~
```

Règle C-A.3 – Accord en nombre. Singulier.

```
1  if ( ^coref(#1[sempl],#2)
2       & ( #2[pl:~,sempl:~]
3           | #2[det,semsg] )
4     ) ~
```

Règle C-A.4 – Accord en nombre. Pluriel.

En règle générale, les noms propres ont un nombre singulier. L'exigence que l'antécédent éliminé ne soit pas un nom propre dans la règle qui rend compte de la reprise par un pronom singulier (C-A.3) n'est cependant pas inutile : dans la phrase suivante le nom propre *les AGF* est pluriel, mais il est repris par un déterminant possessif de nombre sémantique singulier.

(38) De façon générale, les accords conclus entre *les AGF* et *son* nouvel actionnaire à 51 % prévoient que l'Amérique du Nord, l'Europe du Nord et de l'Est, le Royaume-Uni et l'Asie seront sous la responsabilité directe d'Allianz, [...].

10.3.2 Contraintes relationnelles

On appelle « contraintes relationnelles » des règles qui excluent la coréférence entre une expression pronominale e_i et un antécédent potentiel e_j sur la base des relations qu'entretiennent ces expressions entre elles ou avec une ou plusieurs autres expressions de la phrase.

Dans une large mesure, les contraintes spécifiées ici expriment des impossibilités de coréférence qui sont exclues par les contraintes de liage ou de c-commande. On rappelle que ces dernières contraintes sont exprimées sur une représentation spécifique de la structure syntaxique de la phrase sous forme d'un arbre. La structure sous forme d'arbre syntaxique partiel et de dépendances qui nous est donnée en entrée ne permet pas une expression directe des contraintes de liage ou de c-commande (voir la présentation de la théorie du liage et de la c-commande dans la section 6.1).

Expressions reliées à un même verbe

Lorsqu'un pronom clitique non réfléchi est sujet ou complément d'un verbe, il ne peut être coréférent avec une expression qui est elle-même sujet ou complément de ce verbe. Dans la phrase

(39) Il le voit.

les expressions *Il* et *le* ne peuvent avoir la même dénotation.

La règle C-R.1 traduit cette observation.

Si on considère l'exemple (39), les relations que nous voudrions exclure par la règle C-R.1 seraient les deux suivantes :

```
coref(Il,le)
coref(le,Il)
```

Toute hypothèse sur l'existence de la première de ces deux relations est exclue par les règles sur les zones d'antécédence [8]. En revanche, les règles sur les zones d'antécédence n'excluent pas qu'un pronom clitique complément d'un verbe renvoie au sujet de ce même verbe ou à un autre clitique complément du même verbe. Cela est plus particulièrement vrai de la règle Z-PC.9, qui rend compte du cas général pour les pronoms clitiques en disant simplement qu'un pronom renvoie à une expression qui le précède dans la même phrase. Ainsi, pour la phrase suivante,

(40) En réalité, lors de la dernière réunion de la CNDA le 23 mars dernier, six administrateurs avaient persisté dans leur refus de payer, alors que la Commission les menaçait de sanctions disciplinaires.

les règles sur les zones d'antécédence sont telles que le syntagme *la Commission* est considéré comme un antécédent potentiel du pronom *les* (abstraction faite des contraintes d'accord).

La règle C-R.1 ne s'applique donc que pour les reprises par pronom clitique complément. Elle exprime les contraintes suivantes : un pronom clitique `#2`, référentiel, argument d'un verbe `#1`, et une expression dénotante `#3` argument ou sujet du verbe `#1`, ou complément du sujet de `#1` et introduit par *d'entre*, et différente du pronom `#2`, ne peuvent pas être coréférents.

Le lecteur aura remarqué que la règle C-R.1, contrairement aux autres contraintes, ne conclut pas à l'effacement d'une relation `coref`, mais crée une nouvelle relation `non-coref` entre les deux expressions qui instancient `#2` et `#3`. L'effacement d'une éventuelle relation `coref` entre ces deux expressions est effectué par la règle C-R.2. On procède donc ici en deux temps, pour les raisons suivantes.

Étant donné la phrase suivante :

(41) [...] c'est aux journalistes eux-mêmes que revient la responsabilité de défendre les intérêts du public tels qu'ils les imaginent.

[8] La seule règle autorisant qu'un pronom renvoie à une expression qui le suit est la règle Z-PC.5 et cette règle exige que les deux expressions soient reliées à deux verbes différents.

```
1   if ( varg(#1,#2[pron,clit])
2        & ref(#2)
3        & ( varg(#1,#3)
4            | subj(#1,#3)
5            | ( subj(#1,#4)
6                & nmod(#4,?[form:fdentre],#3) ) )
7        & ref(#3)
8        & (#3 ~: #2)
9      )
10  non-coref(#2,#3)
```

RÈGLE C-R.1 – Expressions reliées à un même verbe (1).

```
1   if ( ^coref(#1,#2)
2        & non-coref(#1,#2)
3      ) ~
```

RÈGLE C-R.2 – Expressions reliées à un même verbe (2).

la règle C-R.1, combinée avec la règle C-R.2, exclura la coréférence entre entre le pronom *les* et le pronom *ils*. Cependant, la relation de coréférence est transitive, si bien qu'on voudrait dire non seulement que le pronom *les* ne peut pas être coréférent avec le pronom *ils*, mais aussi qu'il ne peut pas être coréférent avec une expression avec laquelle le pronom *ils* est coréférent.

Pour l'exemple (41), les règles sur les expressions dénotantes et les zones d'antécédence, ainsi que les contraintes d'accord, identifient comme hypothèses de coréférence les relations suivantes :

(i) `coref(ils,journalistes)`
(ii) `coref(ils,intérêts)`
(iii) `coref(les,journalistes)`
(iv) `coref(les,intérêts)`
(v) `coref(les,ils)`

La règle C-R.1 donne lieu à la création de la relation

(vi) `non-coref(les,ils)`

et la règle C-R.2 conduit à l'effacement de la relation (v).

Compte tenu de l'impossibilité que *ils* et *les* aient la même dénotation, l'univers constitué des relations `coref` (i) à (iv) devrait se réduire à une alternative : soit on a les deux relations

(ii) `coref(ils,intérêts)`
(iii) `coref(les,journalistes)`

qui traduisent l'interprétation peu probable selon laquelle ce sont les intérêts du public qui imaginent les journalistes, soit on a les deux relations

(i) `coref(ils,journalistes)`
(iv) `coref(les,intérêts)`

qui traduisent l'interprétation correcte du texte, selon laquelle les journalistes imaginent les intérêts du public.

En d'autres termes, on ne peut avoir à la fois (i) et (ii), ni à la fois (iii) et (iv). Notons que, dans le cas de notre exemple, la sélection de l'une ou l'autre option de l'alternative est déterminée par la résolution d'un des deux pronoms. Par exemple, si on dit que *ils* est coréférent avec *les journalistes*, alors, nécessairement, *les* est coréférent avec *les intérêts du public*. On pourrait imaginer que les contraintes posées par la règle C-R.1 soient mises en mémoire dans le système et que la cohérence des relations de coréférence et non-coréférence soit automatiquement vérifiée, mais ce n'est pas ainsi que fonctionne le système XIP. Toute opération de XIP sur l'univers des relations ne peut résulter que de l'application d'une règle et cette opération est effectuée seulement au moment de l'application de la règle. Aucune règle n'est mémorisée. En créant des relations `non-coref` entre deux expressions par la règle C-R.1, nous pallions cette limite du système en gardant la trace du fait que *ils* et *les* sont reliés au même verbe, qu'à ce titre ils ne peuvent être coréférent et qu'ils ne peuvent donc pas être rattachés au même antécédent. La relation `non-coref` sera utilisée plus tard dans le processus d'analyse pour exprimer cette dernière contrainte, lorsque le rattachement non ambigu d'un des deux pronoms permettra de déterminer quelle hypothèse de coréférence est à exclure pour l'autre.

Exemples. Pour terminer cette présentation de la règle C-R.1, nous donnons ici quelques exemples de son application.

Il est à noter que l'exigence sur le fait que les deux expressions liées par une relation `non-coref` soient des expressions dénotantes exclut les pronoms relatifs, mais la variable `#3` peut néanmoins être instanciée par l'antécédent d'un pronom relatif, puisque s'il existe une relation `subj` ou `varg` entre un relatif `#11` et un verbe `#12`, il existe aussi une relation entre l'antécédent du relatif `#13` et le verbe `#12` (voir l'exemple (42)). Par ailleurs, la règle s'appliquera également pour les verbes coordonnés (exemple (43)), les verbes à l'infinitif (exemple (44)) et les verbes au participe présent (exemple (45)) si une relation `subj` a été identifiée pour ces verbes (voir la définition de la relation `subj` [as-254]).

Dans les exemples suivants, un indice précédé d'une étoile au niveau d'une expression pronominale indique que l'antécédent qui a même indice est exclu par la règle C-R.1. Le verbe sur lequel cette exclusion est fondée est en *italiques*. Dans chacun de ces cas, la règle crée une relation `non-coref` entre l'expression pronominale et l'antécédent exclu.

(42) Officiellement, ce tour de passe-passe autour de la direction des Caisses d'Epargne ne préjuge en rien du [sort]^i qui [lui]$_{*i}$ sera *réservé*, et en particulier au président du directoire, René Barberye.

(43) Selon certains, l'[assureur]^i de Trieste aurait depuis obtenu le feu vert de la Banque d'Italie, mais ne [l']$_{*i}$ a pas pour l'instant *mis* à profit.

(44) La [Réserve]^i fédérale ne devrait pas en effet [la]$_{*i}$ *laisser* franchir la barre des 10 % des dépôts bancaires.

(45) L'avantage évident pour une culture démocratique est que cette [logique]^i sociale dépasse, en [l']$_{*i}$ *intégrant* comme une des structures, une simple distribution des pouvoirs en forme hiérarchique.

Pronom sujet dans une proposition relative et antécédent du relatif

La règle suivante (C-R.3) exprime une contrainte similaire à celle qui est exprimée par la règle C-R.1, mais là où C-R.1 s'applique pour les pronoms clitiques compléments, la règle C-R.3 s'applique pour les pronoms clitiques sujet, dans le contexte d'une proposition relative. En effet, si les règles sur les zones d'antécédence excluent qu'un pronom sujet d'un verbe V_i renvoie à une expression qui le suive et soit complément de V_i, elles n'excluent pas — en particulier la règle Z-PC.9 — que ce pronom renvoie à une expression qui le *précède* et soit complément de V_i. Cela peut se produire lorsque le pronom sujet est le sujet d'une proposition relative.

Pour la phrase suivante,

(46) Paribas s'appuiera pour cela sur le bureau de représentation qu'il a ouvert en 1966.

les règles sur les expressions dénotantes et les zones d'antécédence identifient les deux hypothèses de coréférence suivantes pour le pronom *il* :

```
coref(il,Paribas)
coref(il,bureau)
```

Dans cette phrase le pronom relatif *qu'* et le pronom *il* sont reliés au même verbe et à ce titre ne peuvent être coréférent et, pour ce qui concerne la coréférence, ce qui vaut pour le pronom relatif vaut aussi pour son antécédent : l'expression *le bureau de représentation* et le pronom *il* ne peuvent donc être coréférents.

C'est cette observation que traduit la règle C-R.3, que l'on formule comme suit : un pronom clitique sujet d'un verbe `#4` (l. 2) dans une proposition subordonnée introduite par un pronom relatif `#3` ne peut être coréférent avec l'antécédent du pronom relatif. Dans la règle, le pronom relatif est repéré comme étant une expression qui « connecte » le verbe `#4` (l. 3) [as-261] et qui a un antécédent `#2` (l. 4) [as-262].

```
1  if ( ^coref(#1[pron],#2)
2       & subj(#4,#1)
3       & connect(#4,#3)
4       & antec(#2,#3)
5     ) ~
```

RÈGLE C-R.3 – Pronom sujet et antécédent du relatif.

Il est bon de noter que la contrainte exprimée ici est plus spécifique qu'une simple condition sur la relation à un même verbe du pronom sujet et du pronom relatif dont nous considérons l'antécédent. Dans certains cas, un pronom relatif complément pourra être, plutôt que complément du verbe, complément d'un complément du verbe. Notre règle est formulée de telle manière qu'elle s'applique également dans ce cas. Ainsi, étant donné la phrase :

(47) Seul regret pour Swiss Life, le [CCF]^i, dont [il]$_{*i}$ est actionnaire, n'a pas été retenu pour le CIC, partenaire de bancassurance du GAN.

où le pronom relatif *dont* est complément de *actionnaire* et non du verbe *est*, la règle C-R.3 exclura l'hypothèse selon laquelle le pronom *il* peut être coréférent avec *le CCF*.

Relation entre complément et complété

Outre les relations à un même verbe déjà évoquées, les contraintes relationnelles mettent en jeu trois règles, C-R.4, C-R.5 et C-R.6, présentées page 337, qui tiennent compte de la relation entre le complément d'un syntagme nominal et le syntagme complété.

POSSESSIF DÉTERMINANT UN COMPLÉMENT DE NOM. La règle C-R.4 dit que

– si un déterminant possessif `#1` détermine une expression `#3`

et si

– `#3` est argument d'un syntagme nominal `#2`, argument introduit par une préposition (relation `narg` à trois arguments, l. 4))
– ou `#3` est modifieur d'un syntagme nominal `#2`, modifieur introduit par la préposition *de* (l. 3)

alors `#1` ne peut être coréférent avec `#2`.

L'exigence que l'expression déterminée par le possessif, dans le cas où elle est modifieur, plutôt qu'argument, de l'antécédent potentiel, soit introduite par la préposition *de* est motivée par des exemples tels que les suivants :

(48) Ce qui permet de trouver un dénominateur de gestion du risque commun à l'entité dans sa globalité et cohérent avec l'ensemble de ses activités.

```
1   if ( ^coref(#1[det,poss],#2)
2        & determ(#1,#3)
3        & ( nmod(#2,?[form:fde],#3) |
4            narg(#2,?,#3) )
5       ) ~
```

Règle C-R.4 – Possessif déterminant le complément d'un syntagme nominal.

```
1   if ( ^coref(#1[pron,ton],#2)
2        & ( nmod(#2,?[form:fdentre],#1) |
3            narg(#2,?,#1) )
4       ) ~
```

Règle C-R.5 – Pronom disjoint complément d'un syntagme nominal.

```
1   if ( ^coref(#1,#2)
2        & ( narg(#2,#3[verb,inf])
3            | ( nmod(#2,#6) & adjarg(#6,#3[verb,inf]) ) )
4        & ( ( ( varg(#3,#4)
5                | varg(#3,?,#4)
6                | vmod(#3,?,#4) )
7              & ( determ(#1,#4)
8                  | ( ( narg(#4,?,#5) | nmod(#4,?,#5) )
9                        & determ(#1,#5) ) ) )
10           | varg(#3,#1) )
11      ) ~
```

Règle C-R.6 – Expression pronominale complément dans un syntagme verbal complément de nom.

(49) Sans ce réseau, avec ses nœuds aux droits et obligations clairement identifiés, les pouvoirs (publics ou civils) continueront d'usurper l'autorité démocratique.

En (48), le possessif *sa* est coréférent avec *l'entité*, bien que le syntagme *dans sa globalité* soit modifieur de cette dernière expression. En (49), le possessif *ses* est coréférent avec *ce réseau*, bien que le syntagme *avec ses nœuds* soit modifieur de cette dernière expression.

Nous n'avons pas effectué d'étude systématique visant à identifier de quelle manière tel ou tel type de modifieur ou telle ou telle préposition sont susceptibles d'interdire ou d'autoriser la coréférence dans les cas visés par la règle C-R.4. En l'état actuel de nos connaissances, nous limitons donc notre règle aux cas où le syntagme prépositionnel dont fait partie le possessif est introduit par *de*.

Notons également que les limites importantes de l'analyseur syntaxique en ce qui concerne le rattachement prépositionnel nous ont aussi influencés dans ce choix.

Pronom disjoint complément de nom. La règle C-R.5 est similaire à la précédente, mais concerne les pronoms disjoints qui sont compléments d'un syntagme nominal.

Étant donné un pronom disjoint `#1`, cette règle dit que

- si `#1` est argument d'une expression dénotante `#2` (l. 3)
- ou si `#1` est modifieur d'une expression dénotante `#2`, dans un syntagme prépositionnel introduit par la préposition *d'entre* (l. 2)

alors `#1` ne peut être coréférent avec `#2`.

La condition sur le pronom modifieur exclut que le pronom *eux* et *deux* aient la même dénotation dans la phrase suivante :

(50) Aucun syndicat n'envisage vraiment de rester en dehors des négociations surtout si $[\text{deux}]^i$ d'entre $[\text{eux}]_{*i}$ décident de les reprendre.

La condition sur le pronom argument exclut que, dans la phrase suivante, le pronom *elle* soit coréférent avec *la lutte*.

(51) Et là où existe la propriété privée, la $[\text{lutte}]^i$ contre $[\text{elle}]_{*i}$ est inévitable et l'idée communiste (l'idéologie immortelle des non-propriétaires) surgit naturellement.

L'exemple suivant nous permettra d'illustrer l'exigence qui veut que dans le cas où le pronom modifie un syntagme nominal, il soit introduit par la préposition *d'entre*.

(52) L'expression à elle seule signifie une socialité et une souveraineté.

Pour cette phrase, l'analyseur syntaxique identifie la relation

```
nmod(expression,à,elle).
```

C'est un choix de description de la phrase qu'on pourrait discuter : une alternative serait de noter une relation de modification entre le verbe et le syntagme dont fait partie le pronom, comme on le ferait avec *elle aussi* dans la phrase suivante.

(53) L'expression signifie elle aussi une socialité et une souveraineté.

Comme pour la règle précédente, nous n'avons pas étudié de manière poussée la meilleure représentation possible des syntagmes prépositionnels avec pronom disjoint. Nous limitons notre règle aux cas où la préposition *d'entre* est utilisée, seul cas qui nous semble assurément valide en l'état actuel de nos connaissances.

Expression pronominale complément dans un syntagme verbal complément de nom. Lorsqu'un syntagme verbal infinitif est complément d'un nom (l. 2), comme, par exemple, *pour réduire*, complément de *solutions* dans la phrase :

(54) Toutes les [solutions]^i pour réduire son impact sur [leurs]$_{*i}$ comptes sont les bienvenues.

ou complément d'un adjectif complément d'un nom (l. 3), comme, par exemple, *de les intéresser*, complément de *susceptibles*, qui est complément de *dépouilles*, dans la phrase :

(55) Les acquéreurs potentiels ont pu identifier les [dépouilles]^i susceptibles de [les]$_{*i}$ intéresser.

un pronom clitique complément du verbe à l'infinitif (p. ex. *les* dans (55)) (l. 10), ou un déterminant possessif qui détermine un complément du verbe à l'infinitif ou un complément d'un complément du verbe à l'infinitif (p. ex. *leurs* dans (54)) (l. 4-9), ne peut pas être coréférent avec le nom dont le syntagme verbal est complément.

10.3.3 Contraintes sur les insertions

Les dernières contraintes que nous avons définies mettent en jeu les « insertions » définies au chapitre 7, section 7.3.9. Elles sont exprimées par quatre règles, deux pour les insertions entre parenthèses, deux pour les insertions entre virgules. On ne décrit que les deux règles pour les insertions entre parenthèses (C-I.1 et C-I.3), sachant que les règles pour les insertions entre virgules (C-I.2 et C-I.4) sont exactement parallèles. Les formules XIP sont données page 340.

La règle C-I.1 dit qu'une expression pronominale qui ne figure pas dans une insertion entre parenthèses ne peut avoir pour antécédent une expression qui figure dans une telle insertion.

La règle C-I.3 dit qu'une expression pronominale dans une insertion I_i ne peut avoir pour antécédent une expression qui figure dans une insertion I_j différente de I_i (ce qu'on traduit par le fait qu'il existe une expression qui ne figure pas dans une insertion entre le pronom et son antécédent potentiel).

```
1  if ( ^coref(#1[inser:~],#2[inser])
2      ) ~
```

Règle C-I.1 – Antécédent dans une insertion entre parenthèses (1).

```
1  if ( ^coref(#1[embed:~],#2[embed])
2      ) ~
```

Règle C-I.2 – Antécédent dans une insertion entre virgules (1).

```
1  if ( ^coref(#1[inser],#2[inser])
2       & dans(#3,#2)
3       & dans(#3,#4[inser:~])
4       & (#4 < #1)
5       & (#2 < #4)
6     ) ~
```

Règle C-I.3 – Antécédent dans une insertion entre parenthèses (2).

```
1  if ( ^coref(#1[embed],#2[embed])
2       & dans(#3,#2)
3       & dans(#3,#4[embed:~])
4       & (#4 < #1)
5       & (#2 < #4)
6     ) ~
```

Règle C-I.4 – Antécédent dans une insertion entre virgules (2).

Exemples. L'exemple suivant illustre l'application de la règle C-I.2. Le déterminant possessif *son* ne figure pas dans une insertion ; il ne peut avoir pour antécédent une expression qui figure dans l'insertion entre virgules *beaucoup moins puissant avec un chiffre d'affaires équivalent à celui du GAN*.

(56) Swiss Life, beaucoup moins puissant avec un [chiffre]^i d'affaires équivalent à [celui]^j du [GAN]^k, bénéficie du soutien de [son]$_{*i/*j/*k}$ actionnaire à 25 %, la nouvelle United Bank of Switzerland.

Notons qu'après application des contraintes, un seul antécédent reste possible pour le déterminant possessif *son* : *Swiss Life*.

L'exemple suivant illustre l'application de la règle C-I.3. Les syntagmes *au sens* et *Pierce* dans la première insertion entre parenthèses ne peuvent être antécédents d'aucune des trois expressions pronominales figurant dans la seconde insertion entre parenthèses.

(57) Au besoin trop humain du visible est concédé le culte des reliques et la représentation allusive (indicielle au [sens]^i de [Pierce]^j) du Bouddha (traces de [ses]$_{*i/*j}$ pas, ombrelle [l']$_{*i/*j}$ ayant abrité, arbre de [son]$_{*i/*j}$ Éveil) mais non l'icône.

10.4 Une propriété générale des règles

Pour terminer cette présentation des règles de notre système d'interprétation des expressions pronominales, nous voudrions mettre en avant une de leurs propriétés, qui met en lumière leur spécificité par rapport aux préférences décrites dans le chapitre suivant. Les règles et préférences sont définies de telle manière qu'il est tout à fait possible que, pour certaines expressions pronominales, le système ne propose aucune interprétation. Cette absence de réponse du système ne pourra cependant être que le fait d'un défaut dans les règles (abstraction faite des erreurs dues à l'analyseur syntaxique), pas d'un défaut dans les préférences.

Les trois types de règles peuvent causer une absence d'interprétation pour une expression pronominale e_i :

(i) si l'antécédent de l'expression pronominale e_i n'est pas une expression dénotante en vertu des règles sur les expressions dénotantes ;
(ii) ou si les règles sur les zones d'antécédence spécifient une zone où ne se trouve aucun antécédent pour e_i ;
(iii) ou si les contraintes excluent la coréférence entre e_i et son antécédent ;

alors le système ne donnera pas d'interprétation pour e_i, ou donnera une interprétation erronée. Nous verrons que les préférences, si elles peuvent bien entendu conduire à des erreurs d'interprétation, ne réduisent jamais l'ensemble des antécédents possibles pour une expression pronominale à un ensemble vide. Les préférences ne pourront donc induire en elles-mêmes une absence de réponse pour une reprise.

Chapitre 11

Préférences

Étant donné un texte en entrée, les règles présentées au chapitre précédent définissent pour chaque expression pronominale de ce texte un ensemble d'antécédents possibles. Dans la plupart des cas, cet ensemble sera constitué de plusieurs expressions qui ne sont pas coréférentes entre elles, c'est-à-dire que bon nombre d'ambiguïtés subsistent.

On souhaite cependant que notre système d'interprétation des expressions pronominales puisse fournir une interprétation unique pour chaque expression pronominale du texte analysé. Ce résultat sera obtenu grâce à un ensemble de formules XIP que l'on appelle « préférences ».

La section 11.1 donne une vue générale des préférences et un exemple de leur application. La section 11.2 en fait une description détaillée.

11.1 Vue générale

Les préférences se distinguent des règles présentées au chapitre précédent par le fait qu'elles ne s'appliquent que dans les cas d'ambiguïté, c'est-à-dire lorsqu'il existe pour une même expression pronominale e_i deux relations `coref(`e_i`,`e_j`)` et `coref(`e_i`,`e_k`)` avec e_j et e_k deux expressions différentes. Si, pour une expression pronominale e_i, il n'existe qu'une seule relation `coref` ayant e_i pour premier argument, alors aucune préférence ne s'applique pour cette expression, qui est considérée comme résolue. Les préférences concluent toujours à l'élimination d'un antécédent possible pour une expression pronominale, mais cela à la condition qu'il y ait *au moins deux antécédents possibles* pour cette expression avant application de la préférence. Cette propriété contraste avec la propriété des règles que nous avons évoquée à la fin du chapitre 10 (ces dernières sont susceptibles d'entraîner une absence de réponse, alors que ce n'est pas le cas des préférences).

Une autre raison pour laquelle nous distinguons règles et préférences réside dans le fait que les préférences ont pour nous une valeur beaucoup plus exploratoire que les règles. Comme nous l'avons dit dans la section 6.6, nous ne pourrons

disposer de certains éléments d'information qui semblent nécessaires à l'interprétation des expressions pronominales. Il résulte de ce manque d'information que l'application d'une préférence sur un exemple particulier pourra donner l'impression que, si le résultat est correct, la raison pour laquelle il l'est n'est pas la bonne (voir en particulier la section 11.1.4). Compte tenu de l'information dont nous disposons, les préférences sont le mieux que nous avons pu faire et si elles n'ont sans doute pas une validité absolue, elles ont, pensons-nous, une réelle validité statistique.

11.1.1 Organisation des préférences

Les préférences sont un ensemble de formules XIP, donc un ensemble de formules ordonnées. Dans la plupart des cas, l'ordre des préférences, contrairement aux contraintes, est significatif dans le sens où si l'ordre d'application des préférences était modifié les résultats seraient différents. Le principe qui a guidé l'ordonnancement des préférences est le suivant :

> S'appliquent en priorité les préférences dont on pense qu'elles sont les plus valides, d'après les tests effectués sur le corpus d'étude.

Pertinence de l'ordre des préférences. L'ordre des préférences exprime une hiérarchie des informations utilisées pour la résolution des expressions pronominales, de l'information la plus fiable à la moins fiable. Là où Lappin & Leass ou Mitkov (voir les sections 6.5.3 et 6.5.6, respectivement) expriment la hiérarchie des différents « facteurs » qu'ils utilisent dans leur système de résolution en leur associant des poids différents et en faisant la somme de ces poids au final, nous exprimons cette hiérarchie par l'ordre des préférences.

Les préférences ordonnées de notre système d'interprétation automatique des expressions pronominales seraient exprimables comme un ensemble de préférences pondérées à la Lappin & Leass [1]. À l'inverse, les préférences pondérées de Lappin & Leass ou Mitkov, étant donné les poids utilisés, ne sont pas traduisibles dans un ensemble de préférences ordonnées.

Au départ, le fait d'exprimer les préférences comme un ensemble de formules ordonnées plutôt qu'un ensemble de préférences pondérées résulte d'une contrainte technique : le système XIP n'offre pas la possibilité d'associer un poids aux formules. Les préférences ordonnées présentent cependant un intérêt par rapport aux préférences pondérées, en l'occurrence celui de permettre une évaluation plus fine des résultats, en particulier des échecs, du système. Les préférences implantées dans notre système expriment toujours un choix absolu entre deux antécédents possibles : un antécédent e_i est préféré à un antécédent e_j ou l'antécédent e_j est préféré à l'antécédent e_i. Le fait que le choix soit binaire permet une

[1] Par exemple, l'importance relative de trois préférences ordonnées A, B, C peut être exprimée en associant à A un poids de 100, à B un poids de 10 et à C un poids de 1.

évaluation claire de la préférence : soit le choix est correct, soit il est incorrect. Nous verrons au chapitre 12 que certaines des préférences que nous avons implantées ont une validité quasi absolue sur les deux corpus que nous avons utilisés. Le mérite des préférences ordonnées est de permettre de faire apparaître une telle validité, qui signifie qu'on a défini (ou qu'on pourra définir, avec quelques adaptations) des conditions suffisantes pour déterminer correctement l'interprétation de certains pronoms dans certains contextes.

Par contraste, on notera que dans un système de préférences pondérées qui ne définit pas une hiérachie absolue des préférences tels que les systèmes de Lappin & Leass et Mitkov, l'évaluation ne peut être que globale. À supposer que les poids aient été associés aux préférences de manière optimale, si le résultat final n'est pas parfait, c'est l'ensemble des préférences qui devra être remis en cause.

Non-pertinence occasionnelle de l'ordre des préférences. Si, dans notre système, l'ordre des préférences est le plus souvent pertinent, il n'est cependant pas absolu, c'est-à-dire qu'il pourra être modifié dans certains aspects sans modification des résultats. Par exemple, la préférence 9, décomposée en deux formules, dit globalement qu'un déterminant possessif pouvant être coréférent avec un syntagme nominal ou son complément est de préférence coréférent avec le syntagme complété s'il dénote une personne (9a), avec le complément sinon (9b). La préférence 10, quant à elle, dit qu'un déterminant possessif est de préférence coréférent avec un syntagme dénotant une personne ou avec un pronom. Dans la mesure où la relation complément/complété n'est pas pertinente pour un pronom, l'ordre de ces deux préférences pourrait être inversé sans modification du résultat final.

11.1.2 Résumé de l'information utilisée dans les préférences

Dans la définition des préférences, nous voudrons parfois mettre au même niveau plusieurs types d'information (autrement dit, leur associer le même poids). On fera référence à ces informations différentes dans une même préférence. Une même formule est ainsi susceptible de faire usage d'informations disparates. Par exemple, une de nos préférence dit qu'un pronom clitique datif ou un déterminant possessif renvoie de préférence à un syntagme dénotant une personne (premier type d'information) ou à une expression qui est elle-même une reprise, ce que nous qualifions de « préférence pour la cohésion du discours » (second type d'information).

Nous faisons ici l'inventaire des principaux types d'information utilisés dans les préférences. Entre parenthèses à la suite de chaque type d'information, on indique le numéro identifiant la ou les préférences qui font usage de cette information [2].

[2] Ces numéros renvoient à la description détaillée des préférences (section 11.2). On présente ici les divers types d'information dans une formulation qui, pour être proche d'une formulation

Parallélisme des fonctions. Dans certains contextes, un pronom clitique ayant telle fonction (p. ex. sujet) aura de préférence pour antécédent une expression occupant la même fonction (5 et 9a). Par exemple, dans la phrase suivante :

(1) Si Pierre a un âne, il le bat.

le pronom *il* et son antécédent *Pierre* ont même fonction ; de même pour le pronom *le* et son antécédent *un âne*.

Dans le système de Lappin & Leass (voir section 6.5.3), le parallélisme des fonctions est un des facteurs positifs dans le choix d'un antécédent. Le parallélisme des fonctions a également été utilisé dans certains systèmes de génération automatique de texte pour déterminer des contextes où la pronominalisation d'un syntagme est possible de manière non ambigüe (voir [64]).

Cohésion du discours. On désigne par le terme « cohésion du discours » le fait qu'un locuteur aura tendance à employer, localement, des expressions pronominales de telle manière que celles-ci sont coréférentes entre elles (6, 10 et 12). Dans le texte suivant, les expressions pronominales en *italiques* ont toutes la même dénotation.

(2) Un banquier, comme un chef d'entreprise, ne peut plus se contenter de tenir compte des risques inhérents à *ses* métiers et d'essayer d'optimiser *sa* rentabilité par la provision. Il *lui* faut créer de la valeur, faute de quoi *il* prendrait le risque de mécontenter le marché et de voir *ses* actionnaires déserter. Pour ce faire, *il* dispose de toute une panoplie d'outils de restructuration et d'optimisation de la rentabilité du capital.

La préférence pour la cohésion du discours se retrouve dans la théorie du centrage, où les deux transitions préférées sont celles où le *backward-looking center* d'un énoncé est le même que celui de l'énoncé précédent, le *backward-looking center* d'un énoncé étant par définition désigné par une expression qui est une reprise (voir section 6.4.1). Nous avons vu (voir p. 201) par ailleurs que M. Strube et U. Hahn [84] accordent une place prépondérante au caractère *hearer-old* des entités mentionnées pour déterminer la hiérarchie des *forward-looking centers*. La notion d'entité *heareròld* est plus générale que celle que nous utiliserons (nous nous limitons à une préférence pour une entité déjà désignée par une expression pronominale), mais elle exprime une idée très voisine de notre préférence pour la cohésion du discours.

Lappin & Leass (voir section 6.5.3) expriment également une préférence pour la cohésion du discours en *ajoutant* les poids associés à une expression e_i à celui qui a été associé à la classe d'équivalence dont e_i est déterminée élément. On retrouve également cette préférence chez Günthner et Lehmann [42, p. 149] où les pronoms sont préférés aux syntagmes pleins.

---

de règles, ne vise qu'à donner une première intuition de l'information en question. Les préférences effectivement implantées sont celles qui seront présentées par la suite.

Référence aux personnes. Il semble que, de manière générale, les expressions pronominales soient plus volontiers employées pour faire référence à une personne (9, 10 et 14).

Nous n'avons pas connaissance de systèmes d'interprétation automatique des expressions pronominales qui exprimeraient cette hypothèse. Nous voyons deux raisons possibles à cela. La première est l'hypothèse en question n'est pas pertinente pour l'anglais (langue pour lesquelles la plupart des systèmes sont conçus), puisque le fait qu'un pronom dénote une personne est marqué dans sa forme (*he, she*, par opposition à *it*). La seconde raison pour laquelle la préférence pour une référence aux personnes n'a pas été exprimée auparavant est qu'elle est peut-être dépendante de notre corpus d'étude. En effet, les textes sur lesquels nous avons travaillé parlent essentiellement de sociétés ; en ce sens, la préférence pour une référence aux personnes dont nous faisons l'hypothèse relève peut-être d'une préférence plus générale pour le thème global du discours. Mitkov exprime une préférence qui va dans ce dernier sens avec ses préférences pour les termes du domaine et pour la réitération lexicale (voir section 6.5.6).

Notons que, comme nous évaluerons notre système sur un corpus de même source et de même type que notre corpus d'étude, nous ne serons pas en mesure d'évaluer si la préférence pour une référence aux personnes est spécifique aux corpus utilisés ou si elle a une valeur plus générale (sur la pertinence de l'évaluation que nous proposerons, voir la section 12.5).

Antécédent sujet. Une expression pronominale renvoie de préférence à une expression qui occupe la fonction de sujet (7, 8, 12 et 14).

La préférence pour le sujet a été très souvent exprimée. En ce qui concerne les systèmes d'interprétation automatique, on la retrouve explicitement chez Baldwin et Lappin & Leass (où elle représente le deuxième facteur positif par ordre d'importance, après la préférence pour une expression de la phrase courante) et implicitement chez Mitkov et Hobbs (voir la description des systèmes en question dans la section 6.5). On retrouve également cette préférence dans des systèmes plus anciens, en l'occurrence ceux de Winograd [95], Günthner et Lehmann [42] ou Carter [17] [3].

La préférence pour le sujet se retrouve également dans la théorie du centrage (voir section 6.4.1) où l'entité la plus haute dans la hiérarchie des *forward-looking centers* d'un énoncé est celle qui est dénotée par l'expression sujet.

Proximité. Dans les reprises internes à la phrase, une expression pronominale renvoie de préférence à une expression proche, plutôt qu'à une expression lointaine (6, 8, 11 et 15).

[3] Carter [17, p. 108] remarque qu'« un examen plus poussé révèle des similarités frappantes entre les préférences de Winograd, Hobbs (telles qu'exprimées implicitement par l'algorithme), et Günthner et Lehmann. Toutes expriment des préférences pour des candidats dans les phrases les plus proches et favorisent les sujets par rapport aux objets [...]. »

La préférence pour la proximité n'est exprimée indépendamment d'autres éléments d'information que dans la dernière préférence (15), qui permet d'obtenir une réponse unique pour chaque expression pronominale. On notera que Lappin & Leass détermine le choix ultime entre deux antécédents possibles de la même manière, à cette différence près que nous limitons cette préférence aux reprises internes à la phrase. Pour les reprises externes à la phrase, nous faisons l'hypothèse inverse (voir p. 367).

Autres informations. La liste des différents types d'information présentée ici n'est pas exhaustive. Aux types présentés s'ajoutent diverses références aux fonctions occupées par les expressions ou au fait que les expressions soient ou non complétées par une apposition. Par ailleurs, une préférence peut ne s'appliquer que pour un type d'expression pronominale particulier, par exemple, pour les reprises par pronom disjoint seulement (2).

11.1.3 Formulation succincte des préférences

La présente section décrit, dans une formulation succincte [4], l'ensemble des préférences dans l'ordre de leur application. À chaque préférence (ou dans certains cas, ensemble de préférences proches par le type d'information qu'elles mettent en jeu) est associé un numéro d'ordre et un bref intitulé qui résume l'idée principale de la préférence ou le type d'expression pour lesquelles elle s'applique.

1. Hiérarchie des expressions pronominales. Un pronom ne renvoie pas à un déterminant possessif ou un pronom disjoint s'il peut renvoyer à une expression qui n'est ni un déterminant possessif, ni un pronom disjoint.

2. Pronoms disjoints redondants. Un pronom disjoint complément non essentiel d'un verbe renvoie de préférence au sujet de ce verbe.

3. Expressions dénotant des dates. Une expression pronominale n'est de préférence pas coréférente avec une expression dont le noyau est un nom de date et qui n'est pas sujet.

4. Compléments de lieu. Une expression pronominale n'est de préférence pas coréférente avec un nom propre de lieu introduit par la préposition *à* ou *en* ou un syntagme prépositionnel complément non essentiel d'un verbe et introduit par une préposition pouvant avoir une valeur locative.

5. Parallélisme des fonctions. Dans le contexte d'une coordination de propositions ou d'une exclusion de coréférence entre deux expressions pronominales, un pronom clitique objet renvoie de préférence à un syntagme objet.

[4] Formulation parfois aussi approximative : la formulation complète des préférences est l'objet de la section 11.2.

6. Délimitation par le sujet et cohésion. Une expression pronominale qui peut être coréférente avec un pronom clitique sujet qui est une reprise n'est de préférence pas coréférente avec une expression qui précède ledit sujet.

7. Préférence pour le sujet dans les reprises inter-phrases. Si une expression pronominale dans une phrase p_i peut être coréférente avec une expression sujet dans une phrase p_j différente de p_i, alors l'expression pronominale est de préférence coréférente avec cette expression.

8. Préférence pour le sujet et la proximité dans les reprises intra-phrases. À l'intérieur d'une même phrase, si une expression pronominale peut être coréférente avec une expression e_i sujet ou complément d'un sujet, elle n'est de préférence pas coréférente avec une expression qui précède l'expression e_i, sauf si celle-ci est elle-même sujet.

9. Rapport entre syntagme complément et syntagme complété (possessifs). Un déterminant possessif pouvant être coréférent avec un syntagme nominal ou son complément est de préférence coréférent avec le syntagme complété s'il dénote une personne, avec le complément sinon.

10. Préférence pour une personne et cohésion (datifs et possessifs). Un déterminant possessif ou un pronom clitique datif renvoie de préférence à un syntagme dénotant une personne ou à un pronom.

11. Préférence pour une interprétation locale des possessifs. Les déterminants possessifs sont de préférence interprétés localement, dans une zone délimitée par une expression sujet.

12. Possessifs et contrôle de l'infinitif. S'il peut être coréférent avec une expression qui contrôle le sujet d'un infinitif qui le précède, un déterminant possessif est de préférence coréférent avec cette expression.

13. Cohésion. Une expression pronominale est de préférence coréférente avec une expression qui est elle-même une reprise.

14. Personnes, sujets, syntagmes modifiés par une apposition. Une expression pronominale est de préférence coréférente avec un syntagme dénotant une personne, sinon avec un syntagme sujet, sinon avec un syntagme introduit par la préposition *par* (c'est-à-dire dans la très grande majorité des cas un syntagme qui dénote un agent) ou modifié par une apposition.

15. Préférences finales. Une expression pronominale est de préférence coréférente avec l'antécédent le plus proche si celui-ci est dans la même phrase, avec l'antécédent le plus lointain si celui-ci est dans la phrase précédente.

11.1.4 Exemples

Pour permettre au lecteur de cerner le fonctionnement général des préférences, on donne ici deux exemples de leur application.

Étant donné la phrase suivante,

(3) Contrairement à sa filiale bancaire, le CIC, la privatisation du GAN n'aura pas provoqué de désistements surprises au jour de la remise des offres fermes.

le système de règles présenté au chapitre 10 conduira à l'identification de deux antécédents possibles pour le déterminant possessif *sa* : *la privatisation* d'une part, *le GAN* d'autre part.

Confronté à une telle alternative, un lecteur humain choisira sans hésiter l'antécédent *le GAN* et justifiera probablement ce choix par le fait que si quelque chose est une *filiale*, elle est filiale *d'une société*. On imagine mal ce que pourrait bien être la filiale d'une privatisation. Cependant, nous ne disposons pas, pour l'implantation de notre système de résolution, de l'information selon laquelle qui dit « filiale » dit « filiale de société » et l'exemple (3) illustre bien le fait que bon nombre de cas seront résolus, d'une certaine façon, *par la bande*. Ainsi l'ambiguïté supposée en (3) sera résolue par application d'une préférence (numérotée 9a) qui dit qu'un déterminant possessif, s'il peut renvoyer à un syntagme SN_i ou à son complément SN_j, renverra de préférence à SN_j, sauf si SN_i dénote une personne.

Autre exemple. La phrase suivante contient deux expressions pronominales : le pronom *l'* et le déterminant possessif *son*.

(4) Dans ce contexte, le ministre des Finances, Dominique Strauss-Kahn, vient d'écrire à Alain Le Ray, président du conseil de surveillance du Cencep, pour l'informer de son accord pour une prorogation « à titre exceptionnel » des mandats des instances dirigeantes du groupe.

Pour ces deux expressions, le système de règles défini au chapitre 10 identifie les ensembles d'antécédents suivants :

$A_{\text{l'}}$ = {contexte, Alain Le Ray,}
A_{son} = {contexte, ministre, Alain Le Ray, l'}

Notons que la coréférence entre *l'* et *ministre* est exclue, car cette expression est identifiée comme sujet du verbe *informer* (voir la règle C-R.1).

L'expression *contexte* est exclue des deux ensembles d'antécédents car elle est introduite par la préposition *dans* qui est une préposition à valeur locative (préférence 4). Le pronom *l'* est alors résolu sans ambiguïté.

Les expressions *Alain Le Ray* et *l'* sont exclues de l'ensemble des antécédents de *son* par la préférence 12 qui dit que si un possessif qui suit un verbe à l'infinitif peut être coréférent avec l'expression qui contrôle le sujet de cet infinitif, alors, il est de préférence coréférent avec ladite expression.

Le résultat obtenu est alors correct : le possessif *son* renvoie à *ministre* et le pronom *l'* à *Alain Le Ray*.

Cet exemple, comme beaucoup d'autres qui seront présentés dans ce chapitre, pourra donner au lecteur l'impression que le système a été conçu pour traiter précisément ces exemples et seulement eux. C'est sans doute en partie vrai, puisque nous avons travaillé à partir d'un corpus par définition fini. Celui-ci ne se réduit cependant pas aux seuls exemples présentés dans la thèse. C'est l'évaluation finale qui nous dira si notre système d'hypothèses n'est en fait qu'un système *ad hoc* pour notre seul corpus d'étude.

11.2 Description détaillée

Les préférences implantées dans notre système d'interprétation des expressions pronominales sont ici décrites en détail, dans l'ordre de leur application.

Les formules contiennent systématiquement dans les conditions une référence à une relation `coref(`e_i`,`e_j`)` et une référence à une relation `coref(`e_i`,`e_k`)`. L'une des deux relations est préfixée du symbole « `^` » et la conclusion de la règle (symbole « `~` ») signifie que cette relation doit être effacée (voir la section 8.5.2). Le marquage d'une des deux relations par « `^` » implique dans XIP que les deux relations `coref` auxquelles il est fait référence sont nécessairement différentes. Cette différence, dans nos formules, est une différence dans le second argument de la relation, c'est-à-dire une différence entre deux antécédents possibles (e_j et e_k) pour l'expression qui instancie le premier argument de l'une et l'autre des deux relations (c'est-à-dire l'expression pronominale ambiguë.).

À chaque préférence est associé un numéro d'ordre. Certaines préférences, qui font usage d'une information très proche, sont regroupées sous un même numéro d'ordre et distinguées entre elles par des lettres (p. ex. la préférence 9 se décompose en fait en 9a et 9b). Pour chaque préférence ou groupe de préférences, un intitulé résume l'idée principale de la préférence ou du groupe de préférences. Par sa concision, celui-ci est souvent approximatif et n'a donc qu'une valeur mnémotechnique ; au-delà de cet intitulé, c'est toute la sémantique des préférences effectivement implantées qui doit être prise en compte. Par ailleurs, dans ce qui suit, nous parlerons parfois, un peu improprement, d'un groupe de préférence n comme de *la* préférence n.

1. Hiérarchie des expressions pronominales. Un pronom ne renvoie pas à un déterminant possessif ou un pronom disjoint s'il peut renvoyer à une expression qui n'est ni un déterminant possessif, ni un pronom disjoint.

```
1  if ( ^coref(#1[pron],#2)
2       & (#2[det] | #2[ton])
3       & coref(#1,#3[det:~,ton:~])
4     ) ~
```

Préf. 1 – Hiérarchie des expressions pronominales.

L'exemple suivant permettra de caractériser l'observation que traduit cette préférence [5] :

(5) Un [banquier]^i, comme un chef d'entreprise, ne peut plus se contenter de tenir compte des risques inhérents à [ses$_1$]^j métiers et d'essayer d'optimiser [sa]^k [rentabilité]^l par la [provision]^m. Il$_1$ [lui]$^n_{i/*j/*k/l/m}$ faut créer de la [valeur]^o, faute de quoi [il$_2$]^p_n prendrait le [risque]^q de mécontenter le [marché]^r et de voir [ses$_2$]$^s_{n/o/p/q}$ actionnaires déserter. Pour ce faire, [il$_3$]$_{n/p/q/r/*s}$ dispose de toute une panoplie d'outils de restructuration et d'optimisation de la rentabilité du capital.

On s'intéresse à deux pronoms clitiques de ce texte : *lui* et *il*$_3$ [6]. En vertu des règles définies au chapitre précédent, chacun de ces deux pronoms serait susceptible d'avoir pour antécédent un déterminant possessif. Avant application de la préférence, on a les relations suivantes pour le pronom *lui* :

```
coref(lui,banquier)
coref(lui,ses1)
coref(lui,sa)
coref(lui,rentabilité)
coref(lui,provision)
```

et pour le pronom *il*$_3$:

```
coref(il3,lui)
coref(il3,il2)
coref(il3,risque)
coref(il3,marché)
coref(il3,ses2)
```

La préférence décrite ici exclut les déterminants possessifs de l'ensemble des antécédents potentiels pour les deux pronoms en question, respectivement. Sont donc éliminées les relations :

```
coref(lui,ses1)
```

[5] Les conventions de notation dans les exemples sont les mêmes qu'au chapitre précédent. Les expressions qui sont des antécédents potentiels sont marquées par une lettre en haut à droite. Les expressions qui sont des reprises sont marquées par une série d'indices en bas à gauche renvoyant chacun à un antécédent possible. Les expressions marquées sont celles pour lesquelles il existe une relation `coref` avant application de la préférence. L'exclusion par la préférence évoquée d'une relation `coref` entre une reprise et un antécédent possible est marquée par une étoile préfixant l'indice correspondant au niveau de la reprise. Par exemple, dans (5), le pronom *lui* est marqué de la lettre n pour indiquer son rôle d'antécédent du pronom *il*$_2$ et de la séquence $i/*j/*k/l/m$ pour indiquer qu'il peut avoir pour antécédent les expressions marquées par i, l ou m (marquées en haut à droite) après application de la préférence, mais pas les expressions marquées par j ou k, celles-ci étant précisément exclues par la préférence.

[6] Le pronom *il*$_2$ n'a qu'un antécédent possible, le pronom *lui*, et la préférence en question ne s'applique pas pour lui.

coref(lui,sa)
coref(il$_3$,ses$_2$)

En l'occurrence, les expressions exclues sont des expressions avec lesquelles les pronoms clitiques sont coréférents, mais, dans la mesure où l'ensemble des antécédents pour chacun des deux pronoms contient toujours au moins une expression avec laquelle le pronom est coréférent, notre critère d'évaluation reste satisfait. En effet, parmi les relations qui restent, on a les relations suivantes :

coref(lui,banquier)
coref(il$_3$,lui)
coref(il$_3$,il$_2$)

et notre hypothèse est que c'est en fonction de ces expressions, plutôt qu'en fonction des déterminants possessifs exclus, que sont interprétés les pronoms *lui* et *il*$_3$. À l'appui de cette hypothèse, on observera que le texte de (5) peut être réduit de manière à ce qu'il ne contienne plus les déterminants possessifs en question, sans que l'interprétation des pronoms en souffre :

(6) Un banquier, comme un chef d'entreprise, ne peut plus se contenter de tenir compte des risques [...]. Il$_1$ lui faut créer de la valeur, faute de quoi il$_2$ prendrait le risque de mécontenter le marché [...]. Pour ce faire, il$_3$ dispose de toute une panoplie d'outils de restructuration et d'optimisation de la rentabilité du capital.

Le plus souvent, la préférence 1 éliminera des antécédents avec lesquels les reprises en question sont coréférentes, mais ce ne sera pas toujours le cas. Dans l'exemple suivant, le pronom *Il* n'est pas coréférent avec le possessif *son*.

(7) Bien que la société ne soit pas encore constituée, la composition de [son]^i [comité]^j exécutif a déjà été arrêtée. [Il]$_{*i/j}$ sera présidé par Vicente Tardio, qui vient de la RAS.

2. Pronoms disjoints redondants. Un pronom disjoint complément non essentiel d'un verbe #4 [as-256] renvoie de préférence au sujet du verbe #4.

```
1  if ( ^coref(#1[ton],#2)
2       & coref(#1,#3)
3       & ( vmod(#4,#1)
4           | vmod(#4,?,#1) )
5       & subj(#4,#3)
6     ) ~
```

Préf. 2 – Pronoms disjoints redondants.

Dans l'exemple (8) page suivante, le pronom *eux*, complément non essentiel de *s'entendre*, est coréférent avec *les Belges*, sujet de *s'entendre*.

(8) « Les [dirigeants]^i néerlandais de Fortis ont eu la sagesse de laisser les [Belges]^j s'entendre (ou non) entre [eux]$_{*i/j}$ », plaisante le numéro un de Fortis, interrogé par La Tribune.

3. Expressions dénotant des dates. Une expression pronominale qui peut être coréférente d'une part avec une expression

- dont le noyau est un nom de date [as-240]
- et qui n'est pas sujet,

d'autre part avec une expression

- dont le noyau n'est pas un nom de date,

est de préférence coréférente avec cette dernière. On vise à éliminer principalement par cette préférence les compléments de temps, tels qu'illustrés dans l'exemple suivant.

Étant donné la phrase,

(9) 300 sociétés locales font faillite chaque [mois]^i depuis le [début]^j de l'[année]^k, mais le [mouvement]^l risque de s'accélérer, puisque le [pays]^m traverse [sa]$_{*i/*j/*k/l/m}$ première période de récession depuis une vingtaine d'années.

les syntagmes noyau *chaque mois*, *le début* et *l'année* sont écartés de l'ensemble des antécédents possibles pour le déterminant possessif *sa*.

```
1  if ( ^coref(#1,#2[time])
2       & coref(#1,#3[time:~])
3       & ~subj(?,#2)
4     ) ~
```

Préf. 3 – Expressions dénotant des dates.

4. Compléments de lieu. Une expression pronominale qui peut être coréférente d'une part avec un nom propre de lieu [as-240] introduit par la préposition *à* ou *en* ou un syntagme prépositionnel introduit par une préposition pouvant avoir une valeur locative [as-247] et modifieur d'un verbe, d'autre part avec une expression qui n'a aucune de ces propriétés, est de préférence coréférente avec cette dernière.

Il est à noter que dans le cas où l'antécédent potentiel est un modifieur verbal introduit par une préposition à valeur locative, on n'exige pas que le noyau du syntagme soit un nom de lieu, si bien que les antécédents exclus peuvent ne pas être des compléments de lieu. Ainsi, le syntagme *son projet* est écarté de l'ensemble des antécédents possibles du pronom *lui* dans l'exemple suivant :

(10) National Mutual et Lend Lease ne fusionneront pas : [Axa]^i doit faire une

croix sur son [projet]^j. Cette opération aurait fait de [lui]$_{i/*j}$ l'actionnaire de contrôle d'un des premiers groupes financiers australiens.

Autre exemple, dans lequel le syntagme nominal *en Espagne* est exclu de l'ensemble des antécédents possibles de *ses*.

(11) La nouvelle [société]^i d'[Allianz]^j en [Espagne]^k sera formée par [ses]$_{i/j/*k}$ propres filiales.

```
1   if ( ^coref(#1,#2)
2        & coref(#1,#3)
3        & ( prepobj(?[form:fen],#2[place,proper])
4            | prepobj(?[form:fa],#2[place,proper])
5            | vmod(?,?[locprep],#2)
6          )
7        & ~prepobj(?[form:fen],#3[place,proper])
8        & ~prepobj(?[form:fa],#3[place,proper])
9        & ~vmod(?,?[locprep],#3)
10     ) ~
```

Préf. 4 – Compléments de lieu.

5. Parallélisme des fonctions. On désigne par le terme « parallélisme des fonctions » le fait qu'un pronom renvoie de préférence à une expression ayant même fonction (p. ex. un pronom sujet renvoie de préférence à un syntagme sujet et un pronom objet renvoie de préférence à un syntagme objet).

Dans [64, p. 94], où le but de l'auteur est de définir un système de génération de textes, le parallélisme est défini comme le lien entre deux expressions occupant la même fonction dans deux phrases consécutives dont les arbres syntaxiques sont isomorphes. Décrire l'isomorphisme de deux strutures syntaxiques pose problème dans XIP, dans la mesure où il serait nécessaire d'énumérer toutes les structures possibles dans autant de formules.

Dans le système de Lappin & Leass [55, p. 543], le parallélisme des fonctions est utilisé dans tous les contextes, c'est-à-dire dès que le pronom et son antécédent potentiel occupent la même fonction, sans considération d'un éventuel isomorphisme des deux structures dans lesquelles les deux expressions apparaissent.

Nous nous situerons d'une certaine manière entre ces deux extrêmes, en faisant l'hypothèse que la préférence pour le parallélisme des fonctions est valide dans deux contextes :

- lorsque deux pronoms dépendent d'un même verbe,
- lorsqu'un pronom clitique apparaît dans le second conjoint d'une coordination de propositions et que son antécédent se trouve dans le premier conjoint.

La première situation est celle que nous avons évoquée avec la règle C-R.1 (voir p. 333). Elle peut être illustrée par la phrase :

(12) Si Pierre a un âne, il le bat.

où le pronom *le* renvoie à *un âne* (objet direct) et le pronom *il* renvoie à *Pierre* (sujet).

Pour cette phrase, avant application des préférences que nous allons décrire, on a les relations `coref` suivantes :

(i) `coref(il,Pierre)`
(ii) `coref(il,âne)`
(iii) `coref(le,Pierre)`
(iv) `coref(le,âne)`

et la relation :

(v) `non-coref(le,il)`

Compte tenu de l'impossibilité d'une coréférence entre *le* et *il*, les quatre relations `coref` correspondent en fait à une alternative. Soit on a les relations (i) et (iv), soit on a les relations (ii) et (iii). Les préférences sur le parallélisme des fonctions nous permettront de résoudre cette alternative [7].

La seconde situation d'application des préférences sur le parallélisme des fonctions est illustrée par l'exemple suivant :

(13) Selon certains, l'assureur de Trieste aurait depuis obtenu le feu vert de la Banque d'Italie, mais ne l'a pas pour l'instant mis à profit.

Cette phrase contient une coordination de propositions. Le pronom *l'* dans le second conjoint renvoie à l'objet direct (*le feu vert*) qui figure dans le premier conjoint.

Les formules qui expriment la préférence pour un parallélisme des fonctions sont au nombre de quatre. Les trois premières, référencées 5a, 5b et 5c, visent à résoudre les pronoms clitiques objets, la dernière (5d) exclut comme antécédent possible pour le pronom sujet l'expression qui a été sélectionnée pour le pronom objet. On résout ainsi l'alternative évoquée ci-dessus.

Les trois formules qui permettent de résoudre les pronoms objet dans les contextes décrits ci-dessus définissent une hiérarchie des fonctions les plus probables pour les antécédents desdits pronoms, à savoir :

5a. Un clitique accusatif renvoie de préférence à un antécédent complément d'objet direct,

[7] Les limites du mécanisme d'application des formules XIP dans la situation décrite ici ont été évoquées au chapitre 10, page 333.

5b. sinon [8], un clitique objet (c'est-à-dire accusatif ou datif) renvoie de préférence à un complément d'objet indirect,

5c. sinon, un clitique objet renvoie de préférence à un complément d'un complément d'objet direct ou indirect.

On ne donne ici que la première de ces préférences, sachant que les deux autres sont semblables, modulo la différence dans la fonction de l'antécédent sélectionné et, éventuellement la catégorie du clitique considéré.

On a un pronom `#1` objet d'un verbe `#5` (l. 3), qui peut être coréférent avec l'une ou l'autre de deux expressions `#2` et `#3` et il existe une relation qui exclut la coréférence entre ce pronom et un pronom qui est lui-même une reprise (l. 4-5) ou bien le verbe `#5` est coordonné à un verbe `#6` (l. 6). Dans cette situation, si l'un des deux antécédents possibles est objet direct et l'autre ne l'est pas, alors on élimine ce dernier de l'ensemble des antécédents possibles.

```
1   if ( ^coref(#1,#2)
2        & coref(#1,#3)
3        & varg(#5,#1)
4        & ( ( non-coref(#1,#4)
5              & coref(#4,?) )
6            | coorditems(#6,#5) )
7
8        & varg(#6,#3)
9        & ~varg(?,#2)
10     ) ~
```

PRÉF. 5a – Parallélisme des fonctions (1).

Pour les exemples (12) et (13), reproduits ici, cette préférence donne lieu aux analyses suivantes :

(12) Si [Pierre]^i a un [âne]^j, [il]$_{i/j}$ [le]$_{*i/j}$ bat.

(13) Selon certains, l'assureur de [Trieste]^i aurait depuis obtenu le [feu]^j vert de la [Banque]^k d'Italie, mais ne [l']$_{*i/j/*k}$ a pas pour l'instant mis à profit.

Dans (12), la relation `coref(le,Pierre)` a été effacée ; autrement dit, *un âne* a été sélectionné comme antécédent de *le*. Dans (13), *le feu vert de la Banque d'Italie* est sélectionné comme antécédent de *l'* [9].

Par les préférences sur la résolution des pronoms objets par parallélisme, l'ensemble des antécédents potentiels pour ces pronoms est le plus souvent réduit

[8] Le terme « sinon » veut dire ici « s'il n'existe aucun antécédent qui satisfasse la préférence précédente ».

[9] La coréférence entre *l'* et *l'assureur de Trieste* est exclue par ailleurs du fait que ce dernier syntagme est sujet du verbe dont dépend *l'* (voir la règle C-R.1).

à un élément. Dans le contexte d'une exclusion de coréférence entre deux reprises, on peut dès lors gérer la transitivité de cette exclusion de coréférence.

Pour l'exemple (12), les relations `coref` sont maintenant réduites à l'ensemble suivant :

(i) `coref(il,Pierre)`
(ii) `coref(il,âne)`
(iv) `coref(le,âne)`

avec, toujours,

(v) `non-coref(le,il)`

Dans la mesure où le pronom *le* est interprété comme coréférent avec *un âne*, et où il ne peut être coréférent avec *il*, *il* ne peut être coréférent avec *un âne*. Cette impossibilité est exprimée par la préférence 5d.

5d. Étant donné deux pronoms `#4` et `#1` entre lesquels la coréférence est exclue (`#4` étant un pronom clitique objet, *a priori* résolu par les préférences précédentes), s'il existe parmi les antécédents possibles de `#1` une expression avec laquelle `#4` n'est pas coréférente et une expression avec laquelle `#4` est coréférente, alors exclure cette dernière de l'ensemble des antécédents possibles de `#1`.

```
1  if ( ^coref(#1,#2)
2       & coref(#1,#3)
3       & non-coref(#4,#1)
4       & ~coref(#4,#3)
5       & coref(#4,#2)
6     ) ~
```

PRÉF. 5d – Parallélisme des fonctions (2).

Reprenons l'exemple que nous avons évoqué page 332 pour illustrer la création d'une relation `non-coref`.

(14) [...] c'est aux journalistes eux-mêmes que revient la responsabilité de défendre les intérêts du public tels qu'ils les imaginent.

Avant application des préférences 5a et 5d, on a les relations suivantes :

(i) `coref(ils,journalistes)`
(ii) `coref(ils,intérêts)`
(iii) `coref(les,journalistes)`
(iv) `coref(les,intérêts)`

La préférence 5a exclut la relation (iii) et la préférence 5d exclut la relation (ii).

6. DÉLIMITATION PAR LE SUJET ET COHÉSION. Si une expression pronominale `#1` peut être coréférente avec un pronom clitique sujet `#2` qui est une reprise et

avec une expression #3 qui précède ce pronom, alors écarter l'expression #3 de l'ensemble des antécédents possibles de #1.

```
1  if ( coref(#1,#2)
2       & subj(?,#2)
3       & coref(#2,?)
4       & ^coref(#1,#3)
5       & (#3 < #2)
6     ) ~
```

PRÉF. 6 – Délimitation par le sujet et cohésion.

Comme cela se produit avec la préférence 1, la préférence 6 élimine le plus souvent de l'ensemble des antécédents potentiels pour une reprise e_i des expressions avec lesquelles e_i est coréférente, tout en gardant un antécédent correct dans cet ensemble. Par exemple, dans la phrase suivante, *les principales banques* est exclu de l'ensemble des antécédents de *elles*$_2$, mais *elles*$_1$ figure dans cet ensemble.

(15) Les principales [banques]^i sud-coréennes, emmenées par la Commercial Bank of Korea, ont indiqué hier qu'[elles$_1$]^j décideraient d'ici la fin du mois de juin du sort qu'[elles$_2$]$_{*i/j}$ réserveront à des dizaines d'entreprises du pays, en les classant selon trois types de catégories : normal, sauvable et non viable.

7. PRÉFÉRENCE POUR LE SUJET DANS LES REPRISES INTER-PHRASES. Si une expression pronominale dans une phrase #3 peut être coréférente avec deux expressions figurant dans une phrase #4 différente de #3, et si l'une de ces deux expressions est sujet et l'autre ne l'est pas, alors préférer comme antécédent potentiel de l'expression pronominale l'expression qui est sujet.

```
1  if ( ^coref(#1,#5)
2       & coref(#1,#2)
3       & dans(#3,#1)
4       & dans(#4,#2)
5       & dans(#4,#5)
6       & (#3 ~: #4)
7       & subj(?,#2)
8       & ~subj(?,#5)
9     ) ~
```

PRÉF. 7 – Préférence pour le sujet dans les reprises inter-phrases.

La préférence pour le sujet est une hypothèse qui a souvent été faite (voir, par exemple, la théorie du centrage, décrite dans la section 6.4.1 ou l'algorithme de

Lappin & Leass, décrit dans la section 6.5.3). D'après nos observations, elle est plus nette lorsqu'il s'agit de reprises inter-phrases [10], d'où la présente préférence.

Dans le texte suivant, *l'équipe dirigeante de Suez* est sélectionné comme antécédent du pronom *Elle* (les hypothèses concurrentes étant, à ce stade de l'analyse, *Suez* et *la meilleure valorisation*).

(16) Pour l'heure, l'[équipe]^i dirigeante de [Suez]^j s'emploie à obtenir la meilleure [valorisation]^k possible de son titre. [Elle]$_{i/*j/*k}$ peut déjà compter sur le substantiel montant de plus-values que le groupe a déjà dégagé depuis janvier sur la vente notamment de Sofinco.

8. Préférence pour le sujet et la proximité dans les reprises intra-phrases. Étant donné une expression pronominale `#1` pouvant être coréférente avec l'une ou l'autre de deux expressions `#2` et `#3`, figurant dans la même phrase que `#1`, si l'une des deux expressions (`#3` dans la formule) est sujet ou complément d'un sujet (l. 6-8), et si l'autre (`#2`) n'est pas sujet et précède `#3`, alors `#1` est de préférence coréférente avec `#3`.

```
1   if ( ^coref(#1,#2)
2        & coref(#1,#3)
3        & dans(#4,#1)
4        & dans(#4,#2)
5        & dans(#4,#3)
6        & ( subj(?,#3)
7            | ( ( nmod(#5,?,#3) | narg(#5,?,#3) )
8                 & subj(?,#5) ) )
9        & ~subj(?,#2)
10       & (#2 < #3)
11     ) ~
```

Préf. 8 – Sujet et proximité dans les reprises intra-phrases.

La préférence 8 retient deux antécédents possibles pour le possessif *ses* : *Le ménage* et *la salariée*. Quatre autres antécédents potentiels sont exclus.

(17) Le [ménage]^i demandait au [juge]^j administratif de [Strasbourg]^k de pouvoir bénéficier d'un [crédit]^l d'impôt en France égal à l'[impôt]^m que la [salariée]^n frontalière aurait dû payer sur [ses]$_{i/*j/*k/*l/*m/n}$ revenus perçus en Allemagne.

9. Rapport entre syntagme complément et syntagme complété dans les reprises par déterminants possessifs. On décrit ici deux préférences (9a et 9b) qui concernent les reprises par déterminant possessif lorsqu'ils peuvent

[10] Cette observation a également été faite par B. Baldwin [6, p. 84].

être coréférent à la fois avec un syntagme nominal noyau SN_i et un syntagme nominal SN_j complément de SN_i.

Dans cette situation, la préférence 9a dit que le déterminant possessif est de préférence coréférent avec le syntagme complément, sauf si le syntagme complété dénote une personne [as-240].

La préférence 9b dit que si le syntagme complété dénote une personne, alors c'est lui qui est préféré par rapport à son complément.

```
1  if ( coref(#1[det,poss],#2[person:~])
2       & coref(#1,#3)
3       & ( nmod(#2,#4[form:fde],#3)
4           | narg(#2,?,#3) )
5     ) ~
```

PRÉF. 9a – Préférence pour le complément.

```
1  if ( ^coref(#1[det,poss],#2)
2       & coref(#1,#3[person])
3       & ( nmod(#3,?,#2) | narg(#3,?,#2) )
4     ) ~
```

PRÉF. 9b – Préférence pour le complété s'il décrit une personne.

Les deux exemples suivants illustrent le premier cas :

(18) Contrairement à $[\text{sa}]_{*i/j}$ filiale bancaire, le CIC, la $[\text{privatisation}]^i$ du $[\text{GAN}]^j$ n'aura pas provoqué de désistements surprises au jour de la remise des offres fermes.

(19) L'$[\text{allocation}]^i$ optimale du $[\text{risque}]^j$ et $[\text{sa}]_{*i/j}$ globalisation nous permet alors d'améliorer la rentabilité des fonds propres.

Les deux exemples suivants illustrent le second cas :

(20) Les discussions entre la chancellerie et les $[\text{greffiers}]^i$ des $[\text{tribunaux}]^j$ de commerce sur l'abaissement du tarif d'accès à $[\text{leur}]_{i/*j}$ serveur Minitel (La Tribune du 23 février), qui avaient démarré au début de l'année sur les chapeaux de roue, piétinent.

(21) L'expansion internationale d'Axa vient de connaître un raté : la $[\text{filiale}]^i$ australienne de l'$[\text{assureur}]^j$ français, National Mutual, vient en effet d'abandonner $[\text{son}]_{i/*j}$ projet de fusionner avec Lend Lease.

La première des deux préférences présentées ici met en jeu une sorte de parallélisme des fonctions : le déterminant possessif peut être analysé comme un complément du syntagme qu'il détermine (*sa filiale*, c'est *la filiale de lui*). Cependant, dans le cas des syntagmes dénotant des personnes, cette préférence pour le parallélisme n'est pas valide.

10. Préférence pour une personne et cohésion (datifs et possessifs). La préférence décrite ici est restreinte aux reprises par pronom clitique datif et par déterminant possessif. Si une telle expression peut être coréférente d'une part avec une expression `#2` qui n'est

- ni un nom propre
- ni un syntagme décrivant une personne,
- ni un pronom clitique (sauf *on* (l. 3) [as-243]),
- ni un numéral (c'est-à-dire un pronom numéral),
- ni un déterminant possessif,

et d'autre part avec une expression `#3` qui est soit

- un nom propre,
- un syntagme décrivant une personne,
- un pronom clitique (sauf *on*),
- un numéral (c'est-à-dire un pronom numéral),

alors cette dernière expression est préférable comme antécédent de l'expression pronominale.

L'exigence sur le fait que `#2` ne soit pas un pronom clitique, un numéral ou un déterminant possessif vise à rendre compte de ce qu'on a appelé la tendance à la « cohésion du discours » (voir la section 11.1.2). Il importe de noter, cependant, que l'existence d'une relation de coréférence entre deux déterminants possessifs n'est pas une condition suffisante pour que la préférence élimine un autre antécédent.

```
1   if ( ^coref(#1,#2)
2        & ( #2[proper:~,person:~,clit:~,num:~,det:~]
3            | #2[pron,clit,indef] )
4        & coref(#1,#3)
5        & ( #1[det,poss] | #1[pron,clit,dat] )
6        & ( #3[proper]
7            | #3[person]
8            | #3[clit,indef:~]
9            | #3[num] )
10     ) ~
```

Préf. 10 – Préférence pour une personne et cohésion (datifs et possessifs)

Dans la phrase suivante, le syntagme *Cette activité* est exclu de l'ensemble des antécédents possibles pour le déterminant *sa*.

(22) Cette [activité]^i a permis au [groupe]^j d'affacturage de maintenir [sa]$_{*i/j}$ position de numéro un.

11. Préférence pour une interprétation locale des possessifs. Étant donné un déterminant possessif #1, et

- deux expressions #2 et #3 avec lesquelles peut être coréférent le déterminant possessif #1, l'expression #3 étant dans la même phrase que le possessif #1,
- une expression #7 qui n'est pas un pronom relatif et est sujet à gauche d'un verbe quelconque (l. 5-6),

si #2 précède l'expression sujet #7 (l. 8) et #3 suit ou est cette expression #7 (l. 7), alors le déterminant possessif #1 est de préférence coréférent avec l'expression #3.

La phrase suivante illustre une application de cette préférence.

(23) Mais surtout, [Paribas]^i juge que l'avenir de la [Comit]^j, considérée comme la plus internationale des banques italiennes avec [sa]$_{*i/j}$ filiale Sudameris, aurait plutôt intérêt à se renforcer à l'étranger.

L'expression *Paribas* est exclue de l'ensemble des antécédents du déterminant *sa* car elle précède le sujet *l'avenir*, alors que *la Comit*, autre antécédent possible, suit ce sujet.

```
1  if ( ^coref(#1[det,poss],#2)
2       & coref(#1,#3)
3       & dans(#4,#1)
4       & dans(#4,#3)
5       & subj(#5,#7[rel:~])
6       & (#7 < #5)
7       & ( (#7 :: #3) | (#7 < #3) )
8       & (#2 < #7)
9     ) ~
```

Préf. 11 – Interprétation locale des possessifs.

12. Possessifs et contrôle de l'infinitif. La préférence 12 exprime une préférence pour l'expression qui contrôle le sujet d'un infinitif dans une reprise par déterminant possessif si celui-ci suit le verbe à l'infinitif. L'information en jeu combine la préférence pour le sujet, dans la mesure où c'est le plus souvent une expression sujet qui contrôle le sujet d'un infinitif et une préférence pour la cohésion du discours, si on analyse les phénomènes de contrôle comme une forme d'anaphore.

Étant donné un déterminant possessif #1 et deux expressions #2 et #3 antécédents possibles de #1, si #3 figure dans la même phrase que #1 (l. 3-4) et est sujet d'un verbe à l'infinitif qui précède #1 (l. 5-6), et si #2 n'est pas sujet d'un verbe à l'infinitif (l. 7), alors le déterminant possessif #1 est de préférence coréférent avec l'expression #3.

L'exemple suivant illustre l'application de cette préférence.

(24) Le [gouvernement]^i attend pour la mi-juillet le rapport de la [commission]^j d'enquête parlementaire sur le fonctionnement de l'[institution]^k consulaire, avant d'annoncer [ses]$_{i/*j/*k}$ intentions de réforme.

Le verbe *annoncer* instancie la variable #5 de la formule. Le syntagme *Le gouvernement*, sujet [11] de ce verbe, est retenu comme antécédent pour le possessif *ses*.

```
1  if ( ^coref(#1[det,poss],#2)
2       & coref(#1,#3)
3       & dans(#4,#1)
4       & dans(#4,#3)
5       & subj(#5[inf],#3)
6       & (#5 < #1)
7       & ~subj(?[inf],#2)
8     ) ~
```

PRÉF. 12 – Possessifs et contrôle de l'infinitif.

13. COHÉSION. Si une expression pronominale peut être coréférente d'une part avec une expression #2 qui n'est pas un pronom et n'est pas une reprise, d'autre part avec une expression #3 qui est une reprise, alors l'expression pronominale est de préférence coréférente avec #3.

Comme c'est le cas avec les préférences 1 et 6, la préférence 13 est susceptible d'exclure un antécédent avec lequel une expression pronominale est coréférente, mais un antécédent correct est supposé être malgré tout présent dans l'ensemble d'antécédents réduit après application de la préférence.

Pour la phrase suivante, le déterminant possessif sa_1 est retenu comme antécédent du déterminant possessif sa_2. Par transitivité, sa_2 reste cependant coréférent avec *Suez-Lyonnaise des Eaux* puisque sa_1 est coréférent avec cette dernière expression. Notons que par ailleurs, *la Générale de Belgique*, qui n'est pas un antécédent correct pour sa_2 est aussi exclu.

(25) [Suez-Lyonnaise des Eaux]^i se dote par la même occasion de quelques atouts supplémentaires pour engager [sa$_1$]^j montée en puissance dans la [Générale de Belgique]^k, avec le rachat des 36,6 % qui ne sont pas encore en [sa$_2$]$_{*i/j/*k}$ possession.

14. PERSONNES, SUJETS, SYNTAGMES MODIFIÉS PAR UNE APPOSITION. Un ensemble de quatre formules exprime une hiérarchie de préférences mettant en jeu le fait qu'une expression dénote une personne, occupe la fonction sujet, soit modifiée par une apposition, ou encore soit introduite par la préposition *par*. Avec cette dernière condition, on vise des expressions qui dénotent la plupart du temps un agent.

[11] Plus précisément : expression qui contrôle le sujet de *annoncer*.

```
1  if ( ^coref(#1,#2[pron:~])
2       & ~coref(#2,?)
3       & coref(#1,#3)
4       & coref(#3,?)
5     ) ~
```

PRÉF. 13 – Cohésion.

Nous ferons référence à cet ensemble de formules comme la préférence 14. N'est reproduite, sous le numéro 14a, que la première des quatre préférences. La sémantique des quatre formules est donnée ci-dessous, dans l'ordre de leur application, avec la numérotation 14a, 14b, 14c et 14d.

```
1  if ( ^coref(#1,#2[proper:~,person:~])
2       & coref(#1,#3)
3       & ( subj(?,#3)
4           | prepobj(?[form:fpar],#3)
5           | nmod[appos](#3,?)
6           | nn(#3,?)
7           | #3[proper]
8           | #3[person]
9         )
10      & ~subj(?,#2)
11      & ~prepobj(?[form:fpar],#2)
12      & ~nmod[appos](#2,?)
13      & ~nn(#2,?)
14    ) ~
```

PRÉF. 14a – Personnes, sujets, syntagmes modifiés par une apposition.

14a. Sont préférés les antécédents potentiels qui ont au moins une des propriétés suivantes, par rapport à ceux qui n'en ont aucune :

- syntagmes décrivant une personne ou noms propres.
- syntagmes sujets,
- syntagmes prépositionnels introduits par *par*,
- syntagmes modifiés par une apposition avec ou sans virgule,

Dans la phrase suivante, le syntagme *le tarif* est modifié par une apposition et est retenu comme antécédent du pronom *l'*.

(26) Fin 1997, la chancellerie avait souhaité mettre rapidement un terme au [tarif]^i, jugé prohibitif, de l'[accès]^j au [Minitel]^k (9,21 francs/minute) pour [l']$_{i/*j/*k}$ aligner sur le palier inférieur (5,57 francs).

Dans la phrase suivante, le syntagme *l'industrie*, d'une part, occupe la fonction

sujet, d'autre part, décrit une personne (au sens de [as-240]). Il est préféré comme antécédent du possessif *son* par rapport au syntagme *cette nécessité*.

(27) Car il y a bien longtemps que l'[industrie]^i a intégré cette [nécessité]^j de pilotage par les fonds propres économiques et dimensionné [son]$_{i/*j}$ capital disponible en fonction de ces impératifs.

14b. Si une ambiguïté persiste, sont préférés les antécédents potentiels qui sont noms propres ou décrivent une personne par rapport à ceux qui ne le sont pas.

(28) Pour la CJCE, l'[objet]^i d'une convention n'est pas de garantir au [contribuable]^j que l'imposition due dans un État ne soit pas supérieure à celle qu'[il]$_{*i/j}$ doit payer dans l'autre.

(29) Ces [décisions]^i, estiment certains à Séoul, pourraient renchérir de l'ordre de 15 points de base les conditions de refinancement des trois [banques]^j, à l'heure où [elles]$_{*i/j}$ recherchent des fonds à l'étranger pour financer le nettoyage de leur bilan.

14c. Si une ambiguïté persiste, sont préférés les antécédents potentiels qui ont au moins une des propriétés suivantes, par rapport à ceux qui n'en ont aucune :

- syntagmes sujets,
- syntagmes prépositionnels introduits par *par*,
- syntagmes modifiés par une apposition avec ou sans virgule,

Cela pour résoudre d'éventuelles ambiguïtés qui ne mettraient en jeu que des expressions qui ont satisfait la préférence précédente.

Dans la phrase suivante, les trois syntagmes dont le noyau est entre crochets décrivent des personnes. Le syntagme *Antoine Berhneim* est retenu comme antécédent du possessif *sa* car il occupe la fonction sujet.

(30) Début mai, [Antoine Berhneim]^i paraissait avoir obtenu le soutien d'[Enrico Cuccia]^j pour être renouvelé à la tête de la [compagnie]^k de Trieste et poursuivre [sa]$_{i/*j/*k}$ stratégie consistant notamment à se développer dans la bancassurance en partenariat avec la Comit.

14d. Si une ambiguïté persiste, sont préférés les antécédents potentiels qui sont sujets par rapport à ceux qui ne le sont pas.

Une ambiguïté persiste dans la phrase suivante entre *la compagnie Suisse de Réassurance* et *l'assureur français AGF*, le syntagme *AGF* dans cette dernière expression étant une apposition. L'expression qui est sujet est retenue comme antécédent.

(31) La [compagnie]^i Suisse de Réassurance a franchi en baisse le seuil des 5 % des droits de vote de l'[assureur]^j français AGF, à la suite de l'apport d'une partie de [ses]$_{*i/j}$ titres à l'offre publique déposée par Allianz.

15. Préférences finales. Après toutes les préférences définies jusqu'à présent, des ambiguïtés peuvent subsister. Or on souhaite que le système puisse donner une interprétation unique pour chaque expression pronominale. On utilise pour cela les deux préférences suivantes, référencées 15a et 15b, qui s'appliquent dans l'ordre où elles sont présentées, respectivement pour les reprises internes à la phrase et externes à la phrase.

```
1  if ( ^coref(#1,#2)
2       & coref(#1,#3)
3       & dans(#4,#1)
4       & dans(#4,#3)
5       & (#2 < #3)
6     ) ~
```

Préf. 15a – Préférence finale. Reprises internes à la phrase.

```
1  if ( ^coref(#1,#2)
2       & coref(#1,#3)
3       & dans(#4,#1)
4       & dans(#5,#3)
5       & (#5 ~: #4)
6       & (#3 < #2)
7     ) ~
```

Préf. 15b – Préférence finale. Reprises externes à la phrase.

15a. Si, parmi les antécédents possibles d'une expression pronominale e_i, se trouve un antécédent qui figure dans la même phrase que e_i, sélectionner l'antécédent qui ne précède aucun des autres antécédents.

Dans la phrase suivante, le syntagme *une situation dualiste* est sélectionné comme antécédent de *elle*.

(32) La [distinction]^i kantienne entre les deux usages de la raison nous place dans une [situation]^j dualiste qui pose beaucoup plus de problèmes qu'[elle]$_{*i/j}$ n'en résout.

L'exemple suivant présente selon nous une réelle ambiguïté, que le système résout en sélectionnant *les groupes de travail* comme antécédent du pronom *ils*. Cette interprétation n'est pas celle que nous avons faite lors du codage manuel du corpus.

(33) Certains [syndicats]^i ont également relevé que les [groupes]^j de travail qui avaient été proposés par l'AFB pour préparer les négociations devraient évoluer en groupes de négociations comme [ils]$_{*i/j}$ le souhaitaient.

15b. Si, parmi les antécédents possibles d'une expression pronominale e_i, se trouve un antécédent qui figure dans une phrase différente de celle où se trouve e_i, sélectionner l'antécédent qui précède tous les autres.

Avec cette préférence, on vise indirectement à sélectionner des expressions qui figurent dans la principale de la phrase qui précède la phrase de l'expression pronominale, les propositions enchâssée ayant plutôt tendance à se trouver en fin de phrase. Dans la phrase suivante, l'expression *Albert Frère* est sélectionnée comme antécédent du pronom *lui*.

(34) « [Albert Frère]^i veut avoir une participation importante et gagner de l'argent », résume un [financier]^j. Pour l'instant, la situation pourrait donc [lui]$_{i/*j}$ convenir.

Chapitre 12

Évaluation

Nous avons présenté dans les chapitres précédents un système d'interprétation automatique des expressions pronominales. Il nous faut maintenant évaluer les hypothèses que nous avons formulées.

Lorsqu'il s'agit d'évaluer un système de résolution des pronoms, le choix des critères d'évaluation est largement ouvert. En témoignent les fréquentes différences qui peuvent être observées dans les évaluations de systèmes proposées à ce jour. Par exemple, Lappin & Leass [55] comptent comme une réponse correcte le fait d'identifier un pronom *it* « pléonastique » [1], alors que Kennedy & Boguraev [53] ne les prennent pas en compte [2]. Mitkov [60] [62] semble également compter comme correcte l'identification d'un pronom *it* non anaphorique [3]. En revanche, Hobbs [45], ou encore Peral *et al.* [68], parmi d'autres, excluent ces occurrences de leur évaluation.

Dans le même ordre d'idée, D. Byron [14], dont le propos est précisément l'inconsistance des évaluations d'un auteur à l'autre, donne l'exemple de deux

[1] « Interestingly, the syntactic-morphological filter reduces the set of possible antecedents to a single NP, or identifies the pronom as pleonastic in 163 of the 475 cases (34 %) that the algorithm resolves correctly.» [55, p. 549] Le terme « pléonastique » dans la terminologie de [55] est équivalent à « sémantiquement vide » ; le pronom *it* dans *it is possible that...* est pléonastique.

[2] « We counted 306 third person anaphoric pronouns ; of these, 231 were correctly resoved [...]. The set of 306 "anaphoric" pronouns excluded 30 occurrences of "expletive" *it* not identified by the expletive patterns [...], as well as 6 occurences of *it* which referred to a VP or propositional constituent. » [53, p. 117] Le terme « explétif » est très probablement équivalent au terme « pléonastique » de [55]

[3] Dans [60], est décrite une évaluation sur un ensemble de 223 pronoms dont « 167 sont non anaphoriques (*it* déictique ou non anaphorique) » [p. 873]. La même évaluation est décrite dans [62, p. 1310] dans les termes suivants : « The evaluation in English [...] included texts from different technical manuals [...] which contained a total of 223 pronouns. The robust approach resolved 200 ananphors correctly. », ce qui implique que l'identification des pronoms *it* en question dans [60] est prise en compte dans l'évaluation. Une erreur dans les chiffres fournis par Mitkov est à envisager dans la mesure où il paraît improbable que près de 75 % (167/223) des pronoms d'un corpus soient non anaphoriques.

systèmes de résolution des pronoms qui ne traitent pas les cas de cataphore et qui sont évalués l'un en faisant purement et simplement abstraction de ces cas, l'autre en comptant comme des erreurs les cas de cataphore non résolus ou mal résolus suite à la limitation du système.

Bon nombre d'options sont donc possibles, pour lesquelles on peut trouver une justification satisfaisante, et cette multiplicité d'options nous incitera à évaluer notre système d'interprétation des expressions pronominales sous différents angles.

L'organisation du chapitre est la suivante. Dans la section 12.1, nous présentons l'ensemble des données de l'évaluation : données de la clé, données en sortie, mesures, situations et prédicats d'évaluation. La section 12.2 présente l'évaluation globale du système, dont les différents composants (règles et préférences) sont analysés dans les sections 12.3 et 12.4.

L'évaluation présentée ici ne répond pas à toutes les questions que soulève l'évaluation d'un système d'hypothèses tel que le nôtre. Nous évoquons ces questions dans la section 12.5, puis nous dressons dans une dernière section le bilan du travail présenté dans la deuxième partie de la thèse et envisageons quelques perspectives.

12.1 Données de l'évaluation

Nous rappelons dans cette section quelles sont les données spécifiées dans la clé (12.1.1) et celles de la réponse (12.1.2), ainsi que les mesures d'évaluation qui seront utilisées (12.1.3). Nous envisagerons différentes situations d'évaluation, décrites dans la section 12.1.4 de manière à avoir une évaluation la plus complète possible. La section 12.1.5 décrit en détail les prédicats d'évaluation relativement aux différentes situations envisagées. Enfin, la dernière section présente une manière de voir les données en termes de référents, plutôt que d'antécédents, angle d'approche qui sera adopté pour l'évaluation des contraintes et préférences.

12.1.1 Données de la clé

Étant donné un texte ou un ensemble de textes T, le système est évalué au regard d'une « clé », constituée par l'annotation manuelle de T. L'annotation spécifie ce qui est attendu en sortie du système. Nous avons déjà présenté au chapitre 5 (section 5.1.3) les données spécifiées par la clé. Cependant, cette présentation se limitait aux données pertinentes pour l'interprétation des règles et préférences décrites par la suite, c'est-à-dire à l'information concernant les seules reprises visées par le système. Pour évaluer notre système sous différents angles, comme nous nous proposons de le faire, il nous faut une clé plus complète, qui décrive toutes les occurrences des expressions pronominales retenues, qu'elles soient ou non des reprises. L'objet de la présente section est de définir cette clé complète.

Catégorisation des expressions pronominales

La clé complète caractérise l'ensemble des occurrences des différentes expressions pronominales retenues. La répartition des expressions pronominales est donnée figure 12.1. Elle reprend les distinctions faites au début du chapitre 5.

```
exp
|→ pron
|   |→ rep
|   |   |→ repv
|   |   |→ repx
|   |       |→ srcm
|   |       |→ ncoref
|   |→ nrep
|→ npron
```

FIG. 12.1 – Catégorisation des expressions pour l'évaluation.

Soit `exp` l'ensemble des expressions du corpus analysé. Soit `pron` l'ensemble constitué des pronoms clitiques sujets, accusatifs ou datifs, des pronoms disjoints et des déterminants possessifs du corpus. Cet ensemble est l'ensemble des expressions pronominales retenues dans la section 5.1.1, p. 170, abstraction faite des conditions qui en font ou non des reprises visées par notre système.

Soit `npron` le complémentaire sur `exp` de l'ensemble `pron`, c'est-à-dire l'ensemble des expressions du texte qui ne sont pas une des formes d'expressions pronominales retenues.

L'ensemble `pron` est partitionné en deux sous-ensembles : l'ensemble des expressions qui sont des reprises (`rep`) et l'ensemble des expressions qui ne sont pas des reprises (`nrep`). L'ensemble `nrep` sera principalement constitué d'occurrences du pronom *il* impersonnel et, plus rarement, de quelques occurrences du pronom *le* (voir les exemples (13) et (14) p. 175).

L'ensemble `rep` est lui-même partitionné en deux sous-ensembles : l'ensemble des reprises visées par notre système (`repv`) et l'ensemble des reprises exclues de notre objectif (`repx`). Ce dernier ensemble comprend d'une part les reprises qui renvoient à une source multiple (`srcm`), d'autre part les reprises qui ne sont pas des reprises avec coréférence ou renvoient à une phrase (`ncoref`). Ces dernières reprises se rencontrent avec le pronom clitique *le* (voir les exemples (9) et (12) p. 173).

Information spécifiée pour les différents types d'expressions

La clé spécifie une information différente pour les différents types d'expressions qui appartiennent à l'ensemble `pron` des expressions pronominales retenues.

On représente cette information en associant à chaque expression pronominale e_i un ensemble E_{e_i}.

Étant donné une expression e_i appartenant à l'ensemble `pron` :

- si e_i est une reprise à source multiple, une reprise renvoyant à une phrase ou une reprise interprétée sans coréférence, $E_{e_i} = \{REPX\}$, où $REPX$ est une constante ;
- si e_i n'est pas une reprise, $E_{e_i} = \emptyset$;
- si e_i est une reprise visée par le système, $E_{e_i} = CC_{e_i} - \{e_i\}$, où CC_{e_i} est la chaîne de coréférence à laquelle appartient e_i.

La clé complète pour un corpus T, notée K_T, est l'ensemble des ensembles définis ici.

À titre d'exemple, pour la phrase suivante,

(1) Si, comme semblent le souhaiter la compagnie$_1$ et le gouvernement, le GAN est peu ou prou vendu au niveau de son actif net, soit autour de 15,5 milliards, l'aventure de la compagnie$_2$ de l'assurance se soldera finalement sans perte.

qui contient deux expressions pronominales, *le* et *son*, la clé spécifie les deux ensembles suivants :

$$E_{\text{le}} = \{REPX\}$$

$$E_{\text{son}} = \{\text{compagnie}_1, \text{GAN}, \text{compagnie}_2\}$$

Reprises et sources spécifiées par la clé

Chaque ensemble E_i de la clé est partitionné en deux sous-ensembles. Pour chaque ensemble E_i, on a l'ensemble $\texttt{pron}(E_i)$ des expressions pronominales de E_i :

$$\texttt{pron}(E_i) = \{e \in E_i \mid e \in \texttt{pron}\}$$

et l'ensemble $\texttt{src}(E_i)$ des sources de E_i :

$$\texttt{src}(E_i) = \{e \in E_i \mid e \notin \texttt{pron}\}$$

12.1.2 Données en sortie

Les données en sortie du système ont été présentées au chapitre 9 (section 9.1.1). On rappelle ici ces données, en les complétant pour prendre en compte les reprises exclues de notre objectif et les expressions pronominales qui ne sont pas des reprises.

À partir de l'ensemble des relations `coref` extraites par le système, on définit l'ensemble R_s des reprises en sortie :

$$R_s = \{e_i \mid \text{il existe } \texttt{coref}(e_i, e_j) \text{ en sortie du système}\}$$

Pour chaque reprise e_i de R_s, on a l'ensemble A_{e_i} des « antécédents » de cette reprise :

$$\forall e_i \in R_s,\ A_{e_i} = \{e_j \mid \text{il existe } \texttt{coref}(e_i, e_j) \text{ en sortie du système}\}$$

Pour certaines des expressions pronominales de la clé (ensemble `pron`), le système n'identifiera pas de relation `coref`, soit avec raison, si l'expression n'est pas une reprise dans la clé, soit à tort, si l'expression en question est une reprise.

Soit NR_s l'ensemble des expressions pronominales de la clé qui n'appartiennent pas à R_s (les expressions pronominales qui ne sont pas des reprises en sortie) :

$$NR_s = \texttt{pron} - R_s$$

On postule que :

$$\forall e \in NR_s,\ A_e = \emptyset$$

12.1.3 Mesures d'évaluation

L'évaluation sera effectuée en termes de rappel et précision. On rappelle ici la caractérisation générale de ces deux mesures, qui sont définies par trois valeurs notées *possible*, *effectif* et *correct*. Ces trois valeurs ont la sémantique suivante :

- *possible* = nombre de réponses que doit produire le système ;
- *effectif* = nombre de réponses effectivement produites par le système ;
- *correct* = nombre de réponses correctes produites par le système.

La valeur *possible* est déterminée par la clé, la valeur *effectif* par la sortie du système, la valeur *correct* par comparaison de la sortie avec la clé.

Étant donné ces trois valeurs, le rappel et la précision sont :

$$\textit{rappel} = \frac{\textit{correct}}{\textit{possible}}$$

$$\textit{précision} = \frac{\textit{correct}}{\textit{effectif}}$$

Nous ne ferons pas usage, pour l'évaluation de notre système, des mesures d'analyse des erreurs présentées au chapitre 3 (section 3.6.2), celles-ci s'avérant peu informatives compte tenu des résultats obtenus.

Nous indiquerons néanmoins les valeurs suivantes :

- *incorrect* = nombre de réponses incorrectes produites le système ;
- *manquant* = nombre de réponses manquantes en sortie du système ;
- *superflu* = nombre de réponses superflues produites par le système.

La manière dont sont caractérisées, dans la situation qui nous occupe, les différentes valeurs présentées ici sera explicitée dans la section 12.1.5 ci-après.

12.1.4 Situations d'évaluation

Pour avoir une évaluation la plus complète possible des résultats produits par notre système, en particulier une évaluation qui puisse être rapportée aux méthodes d'évaluation employées par d'autres auteurs, nous distinguons plusieurs situations d'évaluation. Ces différentes situations d'évaluation sont caractérisées, au départ, par l'ensemble des expressions sur lesquelles elles portent. Nous évaluerons les résultats du système :

- sur l'ensemble des expressions pronominales (`pron`) ;
- sur l'ensemble des expressions qui sont des reprises (`rep`) ;
- sur l'ensemble des expressions qui sont des reprises visées par le système (`repv`).

Nous ferons référence à ces différentes situations d'évaluation par E/`pron`, E/`rep` et E/`repv`, respectivement.

Dans l'évaluation E/`repv`, on vise à évaluer plus spécifiquement l'élément central du système : les règles et préférences. Pour cette raison, l'évaluation ne porte que sur les reprises visées par le système. Cette situation d'évaluation est la situation idéale que nous avons évoquée à la fin de la section 5.1.2 (p. 175) :

> Dans une situation idéale, lorsque nous testerons notre système sur un texte quelconque, nous supposerons que les expressions pronominales de ce texte pour lesquelles notre système ne doit pas spécifier d'interprétation ont été identifiées préalablement comme telles.

Les expressions pronominales pour lesquelles le système ne doit pas donner d'interprétation (ou n'a pas été conçu pour donner une interprétation) sont celles qui appartiennent à l'ensemble `nrep` ou à l'ensemble `repx`, autrement dit, celles qui n'appartiennent pas à l'ensemble `repv`.

Le raisonnement qui préside à l'élaboration de l'évaluation E/`rep` est que l'exclusion des reprises à source multiple et des reprises sans coréférence ou renvoyant à une phrase, si elle est justifiée d'un point de vue technique, est arbitraire du point de vue de la résolution des reprises. À cet égard, l'évaluation E/`repv` sera trompeuse : en réalité, le système sera susceptible de trouver une source incorrecte, ou de ne pas trouver de source, pour des expressions qui sont en fait

des reprises [4]. L'objectif de l'évaluation E/`rep` est de tenir compte de ce type d'erreurs.

Enfin, l'évaluation E/`pron` vise à évaluer le système au regard de l'ensemble des expressions pronominales retenues. Dans cette optique, le point essentiel par rapport à l'évaluation E/`rep` est de porter au crédit du système le fait qu'il identifie qu'une expression pronominale *n'est pas* une reprise. Cette situation d'évaluation est celle qu'adoptent Lappin & Leass [55], par exemple.

12.1.5 Prédicats d'évaluation

Les prédicats d'évaluation déterminent dans quelles conditions la réponse du système pour une expression donnée sera jugée « correcte », « incorrecte », « manquante » ou « superflue ». Ces différents jugements, lorsque portés sur les expressions d'un corpus donné, déterminent des valeurs, notées *correct*, *incorrect*, *manquant* ou *superflu*.

Nous distinguons dans la présente section différents prédicats d'évaluation pour les différents types d'expressions de la clé et/ou de la sortie. Selon la situation d'évaluation adoptée, seront pris en compte tous les prédicats d'évaluation ou seulement certains d'entre eux.

Les différentes valeurs déterminées par les différents jugements présentés ici sont notées par des expressions construites sur le modèle suivant :

$$X^{\texttt{type}}$$

où X peut être C, I, M ou S, respectivement pour *correct*, *incorrect*, *manquant* et *superflu* — ou encore I/S dans un cas particulier où une réponse peut être vue comme incorrecte ou superflue selon l'angle dans lequel on se place — et `type` est un des types d'expression selon la catégorisation de la figure 12.1 (p. 371). Par exemple, la valeur $M^{\texttt{repv}}$ correspond au nombre de réponses jugées manquantes pour les reprises de l'ensemble `repv`.

Le tableau 12.1 recense les différentes valeurs qui seront caractérisées ci-après. La partie haute du tableau reprend la catégorisation de la figure 12.1. Une case vide dans le tableau indique que la notion correspondante (*correct*, etc.) n'est pas pertinente pour le type de reprise considéré. La valeur $I/S^{\texttt{nrep}}$ apparaît deux fois pour des raisons qui seront exposées plus loin. La sémantique associée

[4] Poussé à l'extrême, ce raisonnement voudrait que l'on prenne en compte l'ensemble des expressions pronominales anaphoriques du texte, qu'elles appartiennent à l'ensemble des expressions pronominales que nous avons retenues ou non (p.ex. *celui-ci*, les pronoms numéraux, etc. seraient pris en compte). Cette idée est en particulier défendue par Byron [14]. Nous ne la mettons pas en pratique pour deux raisons : la première est que la tâche d'annotation du corpus en serait nettement alourdie, la seconde est que nous voyons mal quelle justification il y aurait à prendre en compte tous les pronoms, plutôt que toutes les reprises anaphoriques, plutôt que toutes les reprises, plutôt que seulement certains pronoms, comme nous le faisons.

	pron			npron
	rep		nrep	
	repv	repx		
correct	$C^{\texttt{repv}}$	$C^{\texttt{repx}}$	$C^{\texttt{nrep}}$	
incorrect	$I^{\texttt{repv}}$	$I^{\texttt{repx}}$	$I/S^{\texttt{nrep}}$	
manquant	$M^{\texttt{repv}}$	$M^{\texttt{repx}}$		
superflu			$I/S^{\texttt{nrep}}$	$S^{\texttt{npron}}$

TAB. 12.1 – Jugements possibles sur les différents types d'expressions.

aux différentes valeurs de ce tableau est détaillée ici, d'abord pour les expressions qui ne sont pas des reprises, puis pour les expressions qui sont des reprises.

Les expressions qui ne sont pas des reprises sont soit des expressions pronominales retenues (ensemble `nrep`, p. ex. un pronom impersonnel), soit des expressions de l'ensemble `npron`.

- $S^{\texttt{npron}}$

Le système est susceptible d'identifier comme une expression pronominale (c'est-à-dire comme appartenant à `pron`) une expression qui ne l'est pas (c'est-à-dire que cette expression appartient en fait à `npron`). Cela est possible en raison de l'ambiguïté catégorielle de certaines formes, p. ex. la forme *le*, pronom ou déterminant, ou la forme *son*, déterminant possessif ou nom commun. Si le système identifie une source pour une expression qui n'est pas une expression pronominale retenue, la réponse est jugée superflue.

- $I/S^{\texttt{nrep}}$

Le système est susceptible d'identifier comme appartenant à `rep` une expression qui appartient en fait à `nrep` et, par voie de conséquence, d'identifier une source pour cette expression. Par exemple, étant donné la phrase :

(2) Le secrétaire général de la CGT, Louis Viannet, a estimé vendredi que le « diktat » du commissaire européen à la concurrence l'avait « emporté ».

notre système dit que le pronom clitique *l'* renvoie au syntagme *le secrétaire général*. Une réponse de ce type sera jugée superflue lors de l'évaluation E/`rep` et incorrecte lors de l'évaluation E/`pron`.

- $C^{\texttt{nrep}}$

La valeur $C^{\texttt{nrep}}$ ne sera pertinente que pour l'évaluation E/`pron`. Dans cette évaluation, on considère comme une réponse correcte du système le fait qu'une expression pronominale qui n'est pas une reprise soit identifiée comme telle.

Les prédicats d'évaluation pour les expressions qui sont des reprises (ensemble `rep`) peuvent aboutir à l'un des trois jugements suivants : une réponse peut être correcte, incorrecte ou manquante. Ces jugements, sur un ensemble de reprises donné, caractériserons trois valeurs correspondantes que nous décomposons selon le sous-type de reprise considéré (reprise visée, reprise à source multiple ou reprise sans coréférence ou renvoyant à une phrase).

• $C^{\texttt{repv}}$ et $C^{\texttt{repx}}$

Si, étant donné une expression e_i appartenant à l'ensemble `rep`, le système identifie une source avec laquelle e_i est effectivement coréférente, alors la réponse du système est correcte.

Notre système est défini de telle manière qu'il n'est susceptible de donner une réponse correcte que pour une expression appartenant à l'ensemble des reprises visées (`repv`), par conséquent la valeur $C^{\texttt{repx}}$ est toujours nulle.

• $I^{\texttt{repv}}$ et $I^{\texttt{repx}}$

Si, pour une expression e_i appartenant à l'ensemble des reprises, le système identifie une source avec laquelle e_i n'est pas coréférente, alors le système donne une réponse incorrecte pour e_i.

• $M^{\texttt{repv}}$ et $M^{\texttt{repx}}$

Si le système n'identifie aucune source pour une expression e_i qui appartient à `rep` dans la clé, alors la réponse pour e_i est dite manquante. Notons que dans le cas des reprises exclues de notre objectif (`repx`), en particulier les reprises de l'ensemble `ncoref` [5], on préfère *a priori* que la réponse du système soit manquante plutôt qu'incorrecte, une réponse manquante pouvant être vue comme explicitant le fait que le système ne sait pas trouver une source pour cette expression.

Valeurs pour les différentes situations d'évaluation

Les valeurs entrant de le calcul des mesures d'évaluation pour les trois situations d'évaluation que nous avons définies peuvent être maintenant caractérisées.

• E/`repv`

$$\begin{aligned}
\textit{possible} &= |\texttt{repv}| \\
\textit{effectif} &= |R_s \cap \texttt{repv}| \\
\textit{correct} &= C^{\texttt{repv}} \\
\textit{manquant} &= M^{\texttt{repv}} \\
\textit{incorrect} &= I^{\texttt{repv}}
\end{aligned}$$

[5] Ces expressions sont caractérisées par le fait qu'elles entrent comme second argument dans une relation `varg[imperso:+]` avec un verbe. Voir la description de cette relation p. 262, et son utilisation dans la règle ER.7, p. 302.

- E/rep

  $possible = |\texttt{rep}|$
  $effectif = |R_s|$
  $correct = C^{\texttt{repv}}$
  $manquant = M^{\texttt{repv}} + M^{\texttt{repx}}$
  $incorrect = I^{\texttt{repv}} + I^{\texttt{repx}}$
  $superflu = S^{\texttt{npron}} + I/S^{\texttt{nrep}}$

- E/pron

  $possible = |\texttt{pron}|$
  $effectif = |R_s \cup \texttt{nrep}|$
  $correct = C^{\texttt{repv}} + C^{\texttt{nrep}}$
  $manquant = M^{\texttt{repv}} + M^{\texttt{repx}}$
  $incorrect = I^{\texttt{repv}} + I^{\texttt{repx}} + I/S^{\texttt{nrep}}$
  $superflu = S^{\texttt{npron}}$

12.1.6 Antécédents et référents

Un dernier aspect de l'évaluation présentée dans la suite du chapitre est qu'elle vise à évaluer les différents composants du système. Lorsque nous évaluerons ces différents composants, nous serons amenés à évaluer des sorties intermédiaires du système, où plusieurs antécédents possibles pourront être donnés pour une même expression. À ce niveau, pour des raisons qui apparaîtront plus loin, nous passerons d'un raisonnement en terme d'antécédents possibles pour une expression pronominale à un raisonnement en terme de *référents* possibles.

Rappelons d'abord ici l'objectif des règles et préférences (voir le tableau 9.1 p. 288). On a en sortie des règles ou d'une préférence un ensemble de couples (e, A_e) où e est une reprise et A_e est l'ensemble des antécédents possibles pour cette reprise selon le système. L'objectif pour une sortie intermédiaire [6] du système est que pour chaque expression e_i qui est une reprise dans la clé [7] et ayant pour ensemble de référence E_{e_i}, soit caractérisé un ensemble A_{e_i}, tel que :

$$\exists\ a \in A_{e_i},\ a \in E_{e_i} \text{ et } |A_{e_i}| \text{ est le plus petit possible.}$$

Pour mesurer l'apport des contraintes et préférences dans le résultat final, nous seront conduits à évaluer dans quelle mesure elles ont permis de réduire le nombre d'antécédents possibles. Il nous semble cependant plus pertinent de raisonner ici en termes de référents possibles.

Ce point est illustré à travers l'exemple suivant, déjà présenté page 372 :

[6] Plus précisément, en sortie des règles sur les zones d'antécédence ou après application d'une contrainte c_i ou préférence p_i.

[7] On ne tient pas compte ici des expressions pronominales qui ne sont pas des reprises dans la clé. La situation d'évaluation pour l'analyse des règles et préférences sera E/repv.

(1) Si, comme semblent le souhaiter la compagnie$_1$ et le gouvernement, le GAN est peu ou prou vendu au niveau de son actif net, soit autour de 15,5 milliards, l'aventure de la compagnie$_2$ de l'assurance se soldera finalement sans perte.

Pour le possessif *son*, la clé spécifie l'ensemble de référence suivant :

$$E_{\text{son}} = \{\text{compagnie}_1, \text{ GAN, compagnie}_2\}$$

Supposons que, en sortie des règles ou d'une préférence, le système identifie comme antécédents possibles pour cette expression l'ensemble suivant :

$$A'_{\text{son}} = \{\text{compagnie}_1, \text{ GAN, gouvernement}\}$$

Supposons maintenant que l'application d'une contrainte ou d'une préférence réduise l'ensemble A'_{son} à l'ensemble A''_{son} suivant :

$$A''_{\text{son}} = \{\text{GAN, gouvernement}\}$$

En vertu de l'objectif exposé page ci-contre, l'ensemble A''_{son} devrait être jugé plus intéressant que l'ensemble A'_{son} car il est plus petit. Nous voulons considérer que ces deux ensembles sont équivalents.

La contrainte ou préférence qui produit l'ensemble A''_{son} à partir de l'ensemble A'_{son} aura éliminé de l'ensemble des antécédents possibles pour le déterminant possessif *son* une expression avec laquelle il est coréférent, mais, en l'occurrence, cela ne doit être considéré ni comme une erreur (l'ensemble A''_{son} est un ensemble d'antécédents correct pour *son*), ni comme un point à porter au crédit de la contrainte ou préférence.

Nous jugerons les ensembles A'_{son} et A''_{son} comme deux ensembles d'antécédents équivalents pour le déterminant possessif *son* de l'exemple (1) en considérant que chacun des deux ensembles spécifie deux référents possibles pour le déterminant en question.

L'ensemble R_e des référents possibles pour une reprise e en sortie est déterminé de la manière suivante.

Soit AN_s l'ensemble de tous les antécédents possibles en sortie :

$$AN_s = \bigcup_i A_{e_i}$$

Soit RS_s (pour « référents superflus ») l'ensemble des antécédents en sortie qui n'appartiennent pas à une chaîne de coréférence CC_i dans la clé (que celle-ci contienne ou non une expression pronominale) :

$$RS = AN_s - \bigcup_i CC_i$$

Pour chaque reprise e en sortie, on définit l'ensemble R_e des « référents possibles » pour e.

$$R_e = \bigcup_j \{A_e \cap CC_j\} \cup \{RS \cap A_e\}$$

Pour une reprise e, les éléments de l'ensemble R_e constituent chacun un sous-ensemble des expressions de A_e, sous-ensemble tel que les expressions qu'il contient sont coréférentes entre elles.

Pour les deux ensembles A'_{son} et A''_{son} ci-dessus, on a les deux ensembles R'_{son} et R''_{son} correspondants :

$$R'_{\text{son}} = \{\{\text{GAN}\}, \{\text{gouvernement}\}\}$$

$$R''_{\text{son}} = \{\{\text{compagnie}_1, \text{GAN}\}, \{\text{gouvernement}\}\}$$

L'objectif pour une sortie intermédiaire (voir p. 378) devient :

$$\exists\ r \in R_{e_i},\ r \subset E_{e_i} \text{ et } |R_{e_i}| \text{ est le plus petit possible.}$$

Les deux ensembles R'_{son} et R''_{son} ont la même cardinalité. Aucune des deux réponses n'est meilleure que l'autre et l'élimination de l'expression *compagnie*$_1$ de l'ensemble des antécédents possibles pour *son* n'est prise en compte, dans l'évaluation, ni comme une erreur, ni comme une bonne réponse de la contrainte ou préférence qui en est la cause (autrement dit, on n'en tient tout simplement pas compte). Pour en tenir compte de manière appropriée, il faudrait se donner un jugement sur la « qualité relative » de deux antécédents corrects pour une expression pronominale, chose que nous avons exclue dès la formulation de nos objectifs (voir la section 5.1.4).

12.2 Évaluation globale

Sont donnés ici les résultats obtenus en sortie finale du système pour les deux corpus *La Tribune*, à savoir le corpus d'étude et le corpus d'évaluation.

L'objectif est de relier chaque reprise de la clé à une source. Pour une expression e_i qui est une reprise en sortie, la réponse du système est correcte si :

(a) $A_{e_i} \subset \texttt{src}(E_{e_i})$ si A_{e_i} et E_{e_i} sont non vides,
(b) $A_{e_i} = E_{e_i}$, sinon.

Rappels : A_{e_i} est l'ensemble des antécédents de e_i en sortie (voir section 12.1.2) ; $\texttt{src}(E_{e_i})$ est l'ensemble des sources de l'ensemble de référence spécifié par la clé pour l'expression e_i (voir section 12.1.1) ; enfin, le critère d'évaluation (b) n'est pris en compte que dans la situation d'évaluation sur l'ensemble des expressions pronominales (E/`pron`).

	nom	acc	dat	clit	poss	ton	TOTAL
nrep	61	1		62			62
(srcm)					(10)		(10)
(ncoref)		(9)		(9)			(9)
repx		9		9	10		19
repv	108	17	22	147	223	18	388
TOTAL	169	27	22	218	233	18	469

TAB. 12.2 – Répartition des expressions pronominales dans le corpus d'étude

	nom	acc	dat	clit	poss	ton	TOTAL
nrep	39			39	3		42
(srcm)	(1)			(1)	(7)		(8)
(ncoref)		(7)		(7)			(7)
repx	1	7		8	7		15
repv	104	10	19	133	216	11	360
TOTAL	144	17	19	180	226	11	417

TAB. 12.3 – Répartition des expressions pronominales dans le corpus d'évaluation

12.2.1 Répartition des expressions pronominales

Les tableaux 12.2 et 12.3 donnent la répartition des pronoms dans le corpus d'étude et le corpus d'évaluation, respectivement. Dans ces tableaux, les lignes correspondent aux catégories définies dans la figure 12.1 p. 371. Les lignes `srcm` et `ncoref` donnent le détail des reprises exclues (ligne `repx`). Les colonnes correspondent aux catégories syntaxiques (telles que définies dans la section 5.1.1 : on distingue les clitiques sujets (`nom`, pour « nominatif »), accusatifs (`acc`) et datifs (`dat`), les déterminants possessifs (`poss`) et les pronoms disjoints (`ton`)). La colonne `clit` donne le nombre total de pronoms clitiques (c'est-à-dire la somme des valeurs des colonnes `nom`, `acc` et `dat`). La colonne TOTAL donne la somme des valeurs des colonnes `clit`, `poss` et `ton`, la ligne TOTAL la somme des valeurs des lignes `nrep`, `repx` et `repv`. La case en bas à droite donne le nombre total d'expressions pronominales.

La répartition des expressions pronominales est très comparable dans les deux corpus. Dans les deux corpus, les déterminants possessifs représentent la plus grande part des expressions pronominales (49,7 % dans le corpus d'étude, 54,2 % dans le corpus d'évaluation), devant les pronoms clitiques sujets (36 % dans le corpus d'étude, 34,5 % dans le corpus d'évaluation). On relève un peu moins d'expressions qui ne sont pas des reprises ou de reprises exclues dans le corpus d'évaluation (10 % et 3,6 %, respectivement, contre 13,2 % et 4 % dans le corpus d'étude). L'influence de ces différences sur les résultats du système ne devrait pas

être significative. Notons cependant la présence dans le corpus d'évaluation de trois déterminants possessifs qui ne sont pas des reprises ; deux de ces déterminants apparaissent dans l'extrait suivant (les possessifs sont en **gras**) :

(3) De fait, si les grandes banques japonaises ont publié des ratios de solvabilité conformes aux normes internationales, seule la Bank of Tokyo Mitsubishi a appliqué la nouvelle réglementation comptable imposant de constater dans le calcul de **son** capital les plus ou moins-values latentes sur **son** portefeuille de long terme.

Rien n'est prévu dans notre système pour identifier ce type d'usage d'une expression pronominale. En l'occurrence, le système dira à tort que ces deux expressions renvoient à *la Bank of Tokyo Mitsubishi*.

La faible quantité de pronoms accusatifs, datifs et disjoints dans les deux corpus sera susceptible de limiter la pertinence de l'évaluation des règles ou préférences qui s'appliquent spécifiquement pour ces formes comme, par exemple, la préférence pour les pronoms disjoints redondants (voir la préférence 2 p. 353) ou la préférence sur le parallélisme des fonctions (voir la préférence 5 p. 356).

Signalons enfin que les deux corpus sont comparables en ce qui concerne la quantité d'expressions pronominales, et surtout de reprises, qu'ils contiennent.

12.2.2 Jugements sur la sortie du système

La comparaison de la sortie du système avec le corpus annoté manuellement (la « clé ») donne lieu à divers jugements sur les différentes reprises notées dans la clé et/ou notées en sortie. Les tableaux 12.4 et 12.5 donnent le nombre de reprises ayant entraîné tel ou tel jugement, suivant le modèle du tableau 12.1 (p. 376). Les valeurs indiquées dans ces tableaux sont celles qui permettront de calculer les mesures d'évaluation. Elles seront commentées dans les sections suivantes.

Chaque ligne correspond à un des quatre jugements possibles. Les colonnes identifient les différents types d'expressions pronominales suivant la catégorisation de la figure 12.1. Un tiret (« – ») dans une case indique que le jugement correspondant n'est pas pertinent pour le type d'expression caractérisé par la colonne.

12.2.3 Mesures d'évaluation en sortie finale

Les tableaux 12.6 et 12.7 donnent les mesures d'évaluation en sortie finale du système pour chacun des deux corpus et dans chacune des situations d'évaluation envisagées dans la section 12.1.4.

Les meilleurs résultats sont obtenus dans la situation E/`repv`, où on ne prend en compte que les reprises visées par le système. La situation d'évaluation la plus sévère est celle où on prend en compte toutes les reprises (E/`rep`). Le fait de porter au crédit du système l'identification correcte des pronoms impersonnels

	repv	repx	nrep	npron	total
correct	348	0	61	–	409
incorrect	30	6	–	–	36
manque	10	13	–	–	23
superflu	–	–	1	0	1
total	388	19	62	0	469

TAB. 12.4 – Jugements sur la sortie du système – Corpus d'étude.

	repv	repx	nrep	npron	total
correct	277	0	35	–	312
incorrect	61	3	–	–	64
manque	22	12	–	–	34
superflu	–	–	7	1	8
total	360	15	42	1	418

TAB. 12.5 – Jugements sur la sortie du système – Corpus d'évaluation.

	E/repv	E/repx	E/pron
possible	388	407	469
effectif	378	385	446
rappel	0,9	0,86	0,87
précision	0,92	0,90	0,92

TAB. 12.6 – Évaluation en sortie finale – Corpus d'étude.

	E/repv	E/repx	E/pron
possible	360	375	417
effectif	338	349	384
rappel	0,77	0,74	0,75
précision	0,82	0,79	0,81

TAB. 12.7 – Évaluation en sortie finale – Corpus d'évaluation.

(situation d'évaluation E/`pron`) améliore bien évidemment les résultats, relativement à la situation E/`rep`, puisque les erreurs consistant à trouver une source pour une expression qui n'est pas une reprise sont déjà prises en compte comme des réponses superflues dans cette dernière situation.

Dégradation des résultats. On constate une nette dégradation des résultats entre le corpus d'étude et le corpus d'évaluation, de l'ordre de 12 points pour le rappel et 10 points pour la précision. Celle-ci n'est pas surprenante. D'une part, un corpus de 388 reprises effectives est manifestement insuffisamment grand pour couvrir tous les phénomènes de reprises pronominales et notre intuition ne pouvait suffire à combler l'écart entre ce corpus et toute l'étendue des reprises possibles. D'autre part, il est normal que les résultats obtenus par un système à base de règles se dégradent sensiblement du corpus d'étude au corpus d'évaluation, dans la mesure où, si on suppose que chaque règle se révèle légèrement moins valide sur le corpus d'évaluation, les erreurs sont alors susceptibles de s'accumuler.

Appréciation globale des résultats. Nous avons dit (section 5.3.2) que notre objectif n'était pas de définir un système d'hypothèses qui ait une validité absolue, mais plutôt un système qui ait une validité statistique, c'est-à-dire que notre objectif était la définition d'une hypothèse valide pour une « quantité importante » de pronoms. À ce titre, notre système rend compte correctement de l'interprétation des trois quarts des expressions pronominales visées. Ces résultats sont comparables à ceux qui ont été obtenus par divers auteurs pour l'anglais [8].

Lappin & Leass [55] annoncent un succès de 86 % sur un corpus de manuels techniques. On notera que deux points dans les conditions d'évaluation sont susceptibles d'influer favorablement sur les résultats de Lappin & Leass, par rapport aux nôtres : (i) l'ensemble des expressions visées inclut les pronoms réfléchis et réciproques[9], (ii) l'analyse syntaxique du texte est corrigée manuellement avant application du système de résolution. Si on fait abstraction des erreurs en sortie de notre système qui sont dues à l'analyseur syntaxique et si on suppose qu'aucune erreur de l'analyseur syntaxique n'est susceptible d'avoir pour effet de pallier un défaut de notre algorithme de résolution, les résultats obtenus par notre système sur l'ensemble des reprises visées sont de 84,7 % pour le rappel et 87,1 % pour la précision sur le corpus d'évaluation (voir p. 387).

L'algorithme « naïf » de Hobbs [45] donne des résultats similaires à ceux de Lappin & Leass (88,3 %) sur un texte qui est aussi parfaitement analysé syntaxiquement et sans prendre en compte les occurrences du pronoms *it* impersonnel ou renvoyant à une proposition.

L'algorithme défini par Kennedy & Boguraev [53] à partir de celui de Lappin & Leass fonctionne avec un taux de succès de 75 %. Comme le font remarquer

[8] Rappelons que nous n'avons pas d'élément de comparaison pour le français.
[9] Les réfléchis et réciproques sont *a priori* plus faciles à traiter.

ces deux auteurs, leur évaluation porte sur des textes *a priori* plus difficiles à analyser (articles de presse), en tout cas plus disparates.

Le système CogNIAC, de Baldwin [7], dans sa version qui donne une interprétation pour tous les pronoms, donne des résultats similaires à ceux obtenus par Kennedy & Boguraev.

12.2.4 Analyse globale des erreurs

La présente section donne une première analyse globale des erreurs avant l'évaluation plus spécifique des règles et préférences dans les sections suivantes. On analyse dans un premier temps les réponses superflues, puis la répartition des erreurs dans la situation E/`repv`.

Réponses superflues

Le système identifie 8 reprises superflues dans le corpus d'évaluation. Une est due à une erreur de désambiguïsation de la forme *l'* comme déterminant. Le système n'étant pas doté de la capacité d'identifier de tels usages, les trois occurrences du déterminant possessif *son* employé dans un sens générique (voir plus haut le tableau 12.3 et l'exemple (3)) donnent toutes lieu à l'identification d'une reprise superflue. Les quatre autres reprises superflues mettent en jeu des pronoms sujet impersonnels qui n'ont pas été identifiés comme tels.

Notons que 35 des 39 pronoms clitiques qui ne sont pas des reprises — tous sont des sujets impersonnels — sont correctement identifiés par l'analyseur syntaxique comme tels, ce qui correspond à un taux de rappel de 89,7 %. La précision de l'analyseur pour cette tâche est de 94,5 % (35/37) ; deux pronoms sont identifiés à tort comme impersonnels.

Répartition des erreurs

On se concentre sur la situation d'évaluation E/`repv` et on appelle « erreurs » l'ensemble des réponses jugée manquantes ou incorrectes. Le système commet 40 erreurs sur le corpus d'étude (30 réponses incorrectes et 10 réponses manquantes) et 83 erreurs sur le corpus d'évaluation (61 réponses incorrectes et 22 réponses manquantes).

On distingue six sources principales d'erreurs, spécifiées dans le tableau 12.8 qui donne la répartition des erreurs dans les deux corpus. Pour chacun des composants du système, on donne le nombre d'erreurs qui lui sont attribuées. La colonne « s » donne le taux de succès des règles et préférences. Le taux de succès pour chacun de ces quatre composants est défini plus loin, au moment de leur analyse détaillée. Avant de commenter ces chiffres, on définit ce que sont les erreurs dues à l'analyseur syntaxique et les erreurs de chaînage.

	étude		*évaluation*	
	nbre	s	nbre	s
analyseur syntaxique	8	–	23	–
expressions dénotantes	2	99,33	1	99,63
zones d'antécédence	8	97,9	19	94,7
contraintes	3	99,23	7	98,06
préférences	17	91,7	23	86,39
chaînage	2	–	10	–
total	40	–	83	–

TAB. 12.8 – Répartition des erreurs dans les deux corpus.

ERREURS DE L'ANALYSEUR SYNTAXIQUE. Notre système d'interprétation des expressions pronominales met en jeu un processus d'analyse syntaxique des textes en entrée, suivi de la partie plus spécifique à l'interprétation des pronoms que nous avons définie et présentée dans la thèse. Il n'est pas toujours facile de dissocier les différentes parties du système pour leur attribuer de façon claire la responsabilité d'une erreur donnée. La convention que nous adoptons est la suivante : on attribue à l'analyseur syntaxique les erreurs en sortie finale qui sont dues à un défaut dans l'arbre syntaxique ou les dépendances spécifiés, ainsi que les erreurs qui sont dues à une absence d'information à valeur sémantique. Lorsque l'information présente au niveau des traits associés à une expression est erronée, l'erreur est attribuée au système global (voir, par exemple, les erreurs d'accord en genre présentées plus loin).

ERREURS DE CHAÎNAGE. Notre objectif en sortie finale du système est que chaque expression pronominale soit reliée à une expression qui ne soit pas elle-même une expression pronominale (une « source »). Cependant, le processus d'analyse dans notre système est tel que ce n'est qu'à la dernière étape que certaines expressions pronominales sont rattachées à une source, par transitivité (voir la section 9.3.5). Cette dernière étape est susceptible de produire ce qu'on appelle des erreurs de « chaînage ».

Considérons le texte suivant, qui contient deux expressions pronominales, dont la source est le syntagme *la banque* :

(4) Enfin, la [banque]^i discute actuellement avec le gouvernement d'un plan de recapitalisation de l'ordre de 2 000 milliards de wons dans les deux prochaines années. Reste à savoir si ces annonces spectaculaires [lui]$_i$ permettront de redorer [son]$_i$ blason international.

La stratégie implantée dans notre système consiste d'une part à relier *lui* à *la banque*, d'autre part à relier *son* à *lui*, puis à la dernière étape, de calculer que, par transitivité, la source de *son* est *la banque*.

En l'occurrence, si notre système rattache correctement *son* à *lui*, il échoue à rattacher *lui* à *la banque* (une erreur dues aux règles sur les zones d'antécédence). Par conséquent, le possessif *son* ne peut pas, au final, être rattaché à *la banque*. On a alors une erreur de chaînage pour cette expression.

Soit R_e la réponse du système pour une expression pronominale e. Si R_e est jugée incorrecte ou manquante en sortie finale du système, mais est jugée correcte [10] en sortie des préférences (c'est-à-dire avant l'étape de transitivité vers les sources), alors la réponse R_e est due à une erreur de chaînage.

Notons que l'étape de transitivité vers les sources est aussi susceptible d'avoir un effet inverse : une réponse incorrecte en sortie des préférences peut donner lieu à une réponse correcte en sortie finale du système, si la transitivité passe par un lien de reprise qui est lui-même incorrect. Le comportement de notre système sur la phrase suivante en donne un exemple :

(5) Pour les opérateurs, la décision chez [Fortis]$_i$ dépend aussi de l'accord [le]$_i$ liant à la [Société Générale de Belgique]$_j$ (SGB) qui s'est irrévocablement engagée à [lui]$_i$ céder [ses]$_j$ titres

Nous intéressent ici le pronom *lui* et le possessif *ses*. En sortie des préférences, notre système relie *lui* à *la Société Générale de Belgique* et *ses* à *lui*. Les deux réponses sont incorrectes, mais, lorsqu'on calcule par transitivité la source de *ses*, alors la réponse pour cette expression devient correcte. Cette situation est la seule de ce type dans notre corpus.

Commentaires. La proportion d'erreurs dues à l'analyseur est plus importante sur le corpus d'évaluation (27,7 % des erreurs) que sur le corpus d'étude (20 %). La principale raison de cette différence de proportion est très probablement due au fait que dans le courant de notre travail sur le corpus d'étude, nous avons été amenés à corriger ou signaler des erreurs dans la grammaire définie pour le français. L'analyseur était par conséquent mieux adapté à notre corpus d'étude.

En supposant qu'aucune erreur de l'analyseur syntaxique n'est susceptible d'avoir pour effet de pallier un défaut de notre algorithme de résolution [11], on peut calculer la borne supérieure du résultat de l'algorithme de résolution en intégrant parmi les réponses correctes les réponses erronées par faute de l'analyseur. Les nouvelles mesures obtenues sont de 84,7 % pour le rappel et 87,1 % pour la précision sur le corpus d'évaluation et de 91,8 % pour le rappel et 93,7 % pour la précision sur le corpus d'étude.

La plus grande proportion d'erreurs dues à l'analyseur fait que l'écart entre les résultats sur le corpus d'étude et les résultats sur le corpus d'évaluation, en ce qui concerne les bornes supérieurs, est moindre que dans la situation d'évaluation réelle.

[10] Voir le critère d'évaluation présenté dans la section 12.1.6.

[11] Ce cas serait celui où l'algorithme produirait une réponse incorrecte si l'entrée syntaxique était correcte et une réponse correcte si l'entrée syntaxique était incorrecte.

Les erreurs de chaînage sont proportionnellement plus nombreuses sur le corpus d'évaluation. La raison en est probablement que le nombre d'erreurs commises avant l'étape de transitivité vers les sources est lui même plus important.

En dehors de l'analyseur syntaxique, les principales sources d'erreur sont les règles sur les zones d'antécédence et les préférences. Alors que le corpus d'évaluation contient un nombre de reprises moins important que le corpus d'étude, tous les composants du système, à l'exception des règles sur les expressions dénotantes, produisent un plus grand nombre d'erreurs. Il y a donc une dégradation des résultats pour chacun de ces composants, approximativement quantifiable comme un doublement du taux d'erreur (le complémentaire du taux de succès).

Nous examinerons plus précisément ces résultats dans les deux sections suivantes, où sont évalués les différents types de règles et les préférences.

12.3 Évaluation des règles

On évalue successivement dans la présente section les trois types de règles : règles sur les expressions dénotantes, règles sur les zones d'antécédence et contraintes. On rappelle que pour l'évaluation des règles et préférences, on n'envisage que la situation E/`repv`, dans la mesure où les erreurs sur les autres types d'expressions pronominales ne sont dues ni aux règles, ni aux préférences.

12.3.1 Règles sur les expressions dénotantes

Dans le corpus d'étude, deux sources ne sont pas caractérisées comme des expressions dénotantes, toutes deux sont des syntagmes nominaux sans déterminant [12] :

(6) Aux termes des statuts mutualistes de la banque et du MRBB, ces personnes sont propriétaires de [**titres**] pour [**leur**] valeur nominale et n'ont aucun droit sur les réserves, même dans le cadre de la démutualisation aux termes du projet élaboré.

(7) Le gouvernement estime que l'accord est « équilibré » car il a reconquis la maîtrise des modalités de [**privatisation**], à condition toutefois qu'elle intervienne d'ici à octobre 1999.

Les éléments qui pourraient expliquer ce qui apparaît ici comme un cas particulier sont selon nous les suivants :

- dans les deux cas, le syntagme nominal source est le plus proche syntagme nominal à droite de l'expression pronominale qui le reprend [13] ;

[12] En gras et entre crochets, l'expression source qui ne satisfait pas les conditions sur les expressions dénotantes et l'expression pronominale qui la reprend.

[13] En considérant *à condition que* comme une locution conjonctive.

– dans les deux cas, le syntagme nominal source est un complément essentiel du nom qu'il complète.

Par ailleurs, dans le second exemple, la privatisation du Crédit Lyonnais est le sujet principal de l'article. Enfin, pour la petite histoire, signalons qu'une étudiante à qui il avait été demandé d'annoter les reprises anaphoriques dans l'article contenant le second exemple a eu la même interprétation que notre système en choisissant *la maîtrise des modalités de privatisation* comme antécédent, ce que nous interprétons comme une indication de la forte tendance à ce que les antécédents des pronoms soient des syntagmes déterminés.

Une expression dénotante n'est pas reconnu comme telle dans le corpus d'évaluation. Il s'agit du syntagme *10 %* dans la phrase suivante, exclu parce que le « nom » *%* a le trait `measure:+`.

(8) L'Office examinera cette année environ 50 000 demandes de marque communautaire, dont [**10 %**] seront rejetées avant [**leur**] publication.

Dans cet emploi, le nom *%* correspond plutôt à ce qu'on a appelé un « nom de fraction » (voir p. 238). Un emploi du nom *%* en tant que mesure se rencontre, par exemple, dans un syntagme tel que *une hausse de 10 %*. Idéalement, une désambiguïsation des emplois de noms de mesures devrait être faite avant application de la règle sur les expressions dénotantes qui les excluent.

Taux de succès. Le corpus d'étude et le corpus d'évaluation contiennent respectivement 297 et 271 sources non pronominales pour 388 et 360 expressions pronominales. On peut mesurer la validité des règles sur les expressions dénotantes (c'est-à-dire leur taux de succès) par le ratio du nombre de sources de la clé qui sont spécifiées comme expressions dénotantes par le système sur le nombre total de sources dans la clé. Pour le corpus d'étude, cette valeur est de 295/297, soit 99,33 % ; pour le corpus d'évaluation, elle est de 270/271, soit 99,63 % [14]. Les propriétés spécifiées par les règles sur les expressions dénotantes sont globalement valides.

12.3.2 Règles sur les zones d'antécédence

L'objectif des règles sur les zones d'antécédence est de construire pour chaque expression pronominale un ensemble d'antécédents possibles qui sera ensuite filtré par les contraintes et les préférences.

Le tableau 12.9 page 391 rappelle sous une forme succincte la sémantique globale des différentes formules définissant les règles sur les zones d'antécédence et le tableau 12.10 page 392 donne le taux de succès de l'application de chacune de ces règles sur le corpus d'étude et sur le corpus d'évaluation. Les lignes du tableau identifient une règle ou une combinaison de règles, puisque plusieurs règles peuvent s'appliquer pour une même expression pronominale.

[14] Ces deux valeurs sont calculées sans tenir compte des erreurs de l'analyseur à ce niveau.

Taux de succès. Les données du tableau 12.10 déterminent le taux de succès des règles sur les zones d'antécédence. La colonne C donne, pour chaque corpus, le nombre d'expressions pronominales pour lesquelles la règle s'applique correctement, c'est-à-dire que l'ensemble d'antécédents défini contient au moins un antécédent correct pour l'expression pronominale en question. La colonne E (pour « erreur ») donne le nombre d'expressions pronominales pour lesquelles la règle s'applique et ne permet pas d'identifier un antécédent correct. La colonne S donne le « taux de succès » des règles, résultat de la division $\frac{C}{C+E}$ en pourcentage.

Les chiffres donnés en regard d'une règle r qui apparaît aussi dans une combinaison de règles correspondent aux cas où seule la règle r s'applique ; par exemple, pour 25 pronoms clitiques, seule la règle Z-PC.8 contribue à la définition de l'ensemble des antécédents possibles et pour 14 autres pronoms clitiques, cet ensemble est défini à la fois par la règle Z-PC.8 et par la règle Z-PC.9.

Dans certains cas, aucune des règles ne permet de définir un ensemble d'antécédents possibles pour une expression pronominale donnée. Ces cas sont recensés dans des lignes intitulées « aucune règle ».

Enfin les chiffres donnés ici sont calculés abstraction faite des erreurs dues aux règles sur les expressions dénotantes ou à l'analyseur syntaxique. Par exemple, pour l'exemple (6) cité plus haut, on suppose que le syntagme *titres* a été identifié comme une expression dénotante et que la règle Z-DP.4 s'applique alors correctement.

Les règles sur les zones d'antécédence ont un taux de succès sur l'ensemble des expressions pronominales visées de 97,9 % pour le corpus d'étude et de 94,7 % pour corpus d'évaluation. Sur les deux corpus, elles donnent des résultats légèrement meilleurs pour les déterminants possessifs que pour les pronoms. La dégradation des résultats du corpus d'étude au corpus d'évaluation est de 3,2 %. Si ce chiffre est en soit peu élevé, il indique cependant que le taux d'erreur de ces règles a plus que doublé (de 2,1 % à 5,3 %).

Comme nous l'avons indiqué dans la présentation des règles sur les zones d'antécédence, les règles qui s'appliquent en premier pour un type d'expression donné visent à rendre compte de cas particuliers. Cet aspect du système se reflète dans les chiffres du tableau 12.10 : les règles Z-PC.8 et Z-PC.9, seules ou combinées, s'appliquent correctement pour 87 % des pronoms clitiques du corpus d'évaluation ; la règle Z-DP.4 s'applique correctement pour 87 % des déterminants possessifs.

Les résultats produits par les règles sur les zones d'antécédence mettent en lumière trois problèmes principaux, qui seront discutés dans les trois sous-sections suivantes :

- un manque de généralité ;
- un problème au niveau de la définition algorithmique du système, qui empêche l'application conjointe des contraintes et des règles sur les zones d'antécédence ;

Pronoms clitiques	
Z-PC.1	tournures interrogatives
Z-PC.2 ou Z-PC.3	sujet d'incise ; antécédent dans la phrase précédente ou la deuxième phrase précédente.
Z-PC.4	antécédent en position topique
Z-PC.5	cataphore
Z-PC.6	pronom entre le sujet et le verbe principal ; antécédent dans la même phrase ou dans la phrase précédente.
Z-PC.7	pronom dépendant du verbe principal, avec enchâssée à gauche ; antécédent dans la même phrase.
Z-PC.8	pronom dépendant du verbe principal, sans enchâssée à gauche ; antécédent dans la phrase précédente
Z-PC.9	cas général ; antécédent dans la même phrase.
Z-PC.8 et Z-PC.9	pronom dépendant du verbe principal, sans enchâssée à gauche ; antécédent dans la phrase précédente ou dans la même phrase.
Pronoms disjoints	
Z-PD.1	pronom modifiant le sujet de la principale ou d'une incise ; antécédent dans la phrase précédente.
Z-PD.2	pronom complément du verbe principal ou précédant le sujet ; antécédent dans la phrase précédente.
Z-PD.3	pronom après le premier sujet ; antécédent dans la même phrase.
Z-PD.2 et Z-PD.3	pronom complément du verbe principal et après le premier sujet ; antécédent dans la phrase précédente ou la même phrase.
Déterminants possessifs	
Z-DP.1	possessif entre le sujet et le verbe ; antécédent dans la phrase précédente.
Z-DP.2	possessif déterminant le sujet principal ; antécédent dans la phrase précédente.
Z-DP.3	cataphore
Z-DP.4	cas général ; antécédent dans la même phrase.
Z-DP.1 et Z-DP.4	possessif déterminant le sujet principal ; antécédent dans la phrase précédente ou dans la même phrase
Z-DP.3 et Z-DP.4	cataphore ou antécédent précédant dans la même phrase.

TAB. 12.9 – Rappel synthétique des règles sur les zones d'antécédence.

règles	*étude*			*évaluation*		
	C	E	S	C	E	S
Z-PC.1	2	0	100	6	0	100
Z-PC.2 ou Z-PC.3	5	0	100	4	0	100
Z-PC.4	1	0	100	1	1	50
Z-PC.5	5	0	100	2	1	66,7
Z-PC.6	0	0	–	0	0	–
Z-PC.7	1	0	100	4	0	100
Z-PC.8	33	1	97	25	0	100
Z-PC.9	69	0	100	66	7	90,4
Z-PC.8 et Z-PC.9	26	1	96,2	14	0	100
aucune règle	–	3	–	–	3	–
total – clitiques	142	5	96,6	122	11	91,7
Z-PD.1	3	0	100	0	0	–
Z-PD.2	3	0	100	1	0	100
Z-PD.3	10	0	100	6	1	85,7
Z-PD.2 et Z-PD.3	2	0	100	3	0	100
total – disjoints	18	0	100	10	1	90,9
Z-DP.1	0	0	–	0	0	–
Z-DP.2	7	0	100	9	0	100
Z-DP.3	6	0	100	9	2	81,8
Z-DP.4	200	2	99	184	4	97,9
Z-DP.1 et Z-DP.4	3	0	100	3	0	100
Z-DP.3 et Z-DP.4	4	0	100	4	0	100
aucune règle	–	1	–	–	1	–
total – possessifs	220	3	98,7	209	7	96,8
total – expressions visées	380	8	97,9	341	19	94,7

TAB. 12.10 – Évaluation des règles sur les zones d'antécédence.

– une sous-estimation générale des cas où une expression renvoie à un antécédent figurant dans la phrase précédente.

Manque de généralité

Les règles sur les zones d'antécédence souffrent d'un manque de généralité : d'une part, on a une règle, la règle Z-PC.6, qui spécifie un contexte tellement particulier qu'il ne se rencontre ni dans le corpus d'étude, ni dans le corpus d'évaluation, d'autre part, on rencontre dans le corpus d'évaluation trois expressions pronominales pour lesquelles aucune hypothèse n'est formulée :

(9) « Le redressement du [**GAN**] nécessitera que nous y investissions la totalité de son cash-flow. Mais une fois qu'[**il**] aura été redressé, nous n'aurons aucune objection à sa cotation », souligne Gilles Laporte.

(10) Au surplus, certains [**donateurs**] n'interrogeront pas l'administration fiscale, et n'utiliseront pas la nouvelle procédure, de peur des conséquences négatives dans l'hypothèse où l'administration ne donnerait pas son accord sur une évaluation. Comment [**les**] rassurer ?

(11) Forte de [**ses**] 47 milliards de fonds propres réévalués en 1997 — dont 28 milliards de francs de fonds propres comptables et 19 milliards de francs de plus-values latentes — [**Groupama**] considère qu'elle a les moyens de ses ambitions.

Ces trois exemples mettent en jeu des structures non prévues par les règles [15] et les règles Z-PC.9 et Z-DP.4, qui traitent le cas général où une expression renvoie à un antécédent dans la même phrase, ne s'appliquent pas car aucune expression dénotante ne précède l'expression pronominale dans la phrase.

Notons que quatre cas de ce type apparaissent également dans le corpus d'étude [16] :

(12) Alors que la Belgique s'émeut de cette perspective, Maurice Lippens estime qu'« il est trop tard pour pleurer sur la [**SGB**]. Il fallait le faire en 1988 quand [**elle**] est passée sous pavillon étranger ».

(13) Pour ne pas entraver au-delà du raisonnable la privatisation, cette [**clause**] sera, tout comme le droit de neutralisation, vendue au(x) futur(s) actionnaire(s). Notons qu'au passage, [**elle**] vaudra plus cher que la clause prévue dans l'ancien modèle.

(14) « Mais c'est aussi un [**outil**] de gestion du temps : sans [**l'**]utiliser comme moyen de tarification, nous essayons de hiérarchiser le temps d'investissement des commerciaux en fonction du capital économique alloué.

[15] Dans le cas de l'exemple (11), la cataphore n'est pas identifiée car la règle exige que le sujet soit précédé d'une virgule.

[16] Rappel : l'exemple (14) contient deux phrases, délimitées par le symbole « : ».

(15) Les [**Generali**] se disent aujourd'hui prêts à soutenir la stratégie de Mediobanca de vouloir réunir Comit et Banca di Roma. Quitte à devoir gérer, en cas de mariage effectif, [**ses**] accords avec la Comit, [...].

On pourrait formuler une règle (appelons-la « règle de nécessité d'un antécédent ») s'appliquant après toutes les autres règles sur les zones d'antécédence et qui dise que si aucun antécédent possible n'a été trouvé pour une expression pronominale, alors les expressions dénotantes de la phrase précédente sont des antécédents possibles, mais le problème évoqué dans le paragraphe suivant montre que cette solution n'est pas satisfaisante.

Problème de l'algorithme d'application des règles

Dans certains cas, les règles sur les zones d'antécédence spécifient pour une expression pronominale un ensemble d'antécédents non vide, mais tous ses éléments sont ensuite éliminés par application des contraintes. On obtient alors un résultat comparable à celui que nous venons d'évoquer pour les exemples (9) à (14) : aucune hypothèse n'est faite en sortie des règles pour certaines expressions pronominales.

L'exemple suivant illustre de manière simple ce problème.

(16) + Le Parlement a adopté début mai une [**loi**] reconnaissant la responsabilité sans faute des fabricants pour les dommages causés par leurs produits.
+ Les assureurs espèrent qu'[**elle**] permettra de redresser les comptes de la branche.

Pour cet exemple, la règle Z-PC.9 s'applique et identifie *les assureurs* comme antécédent possible pour le pronom *elle*. Cet antécédent est le seul possible et il est par la suite exclu par les contraintes d'accord en nombre et genre. Dans la mesure où l'ensemble d'antécédents possibles est non vide, la solution de notre « règle de nécessité d'un antécédent » n'est pas opérationnelle. Il faudrait appliquer cette règle après les contraintes, mais on aurait alors un nouvel ensemble d'antécédents, pour lesquels il faudrait de nouveau appliquer les contraintes, ce qui dans XIP s'exprimerait par une duplication des formules spécifiant les contraintes autour de la nouvelle règle de nécessité d'un antécédent. On atteint ici une limitation certaine du système XIP pour la résolution des expressions pronominales.

Une manière de contourner cette limitation pourrait être de spécifier dans un premier temps des zones d'antécédence assez larges pour garantir un rappel parfait, puis d'appliquer les contraintes, puis de restreindre les zones définies dans un premier temps. Nous verrons que les zones d'antécédence que spécifient nos règles sont de toute manière trop réduites.

Une autre solution serait de compléter le système XIP avec un mécanisme de mise en mémoire de contraintes (une contrainte étant vue comme une formule concluant à l'effacement d'une relation R) de telles sorte que les contraintes

puissent s'appliquer *conjointement* aux règles qui spécifient la création de la relation R. Notons que ce mécanisme pourrait aussi être utilisé pour résoudre le problème des exclusions de coréférence que nous avons évoquées au chapitre 10, page 332 : pour garder la trace d'une exclusion de coréférence entre deux pronoms, la règle C-R.1 crée une relation `non-coref` entre eux, à laquelle il est ensuite plusieurs fois fait référence pour éviter que les deux pronoms renvoient au même antécédent.

À travers un deuxième exemple pour lequel aucun antécédent n'est trouvé pour un pronom, nous voudrions montrer que raisonner en faisant abstraction des contraintes dans la spécification des zones d'antécédence peut avoir son intérêt du point de vue de la personne qui développe le système. Dans l'exemple suivant, le syntagme *Son dépôt* est analysé à tort comme un syntagme détaché en début de phrase et considéré comme le seul antécédent possible pour le pronom *elle* (c'est la règle Z-PC.4 qui s'applique). On a certes là une erreur des règles sur les zones d'antécédence, mais elle est due à une mauvaise spécification de la structure de la phrase, en l'occurrence une absence de prise en compte des propositions participe, non au fait que *Son dépôt* et *elle* ne s'accordent pas.

(17) Une [**société**] anglaise qui commercialise du matériel électronique a voulu déposer la marque « Orange ». Son dépôt rejeté par l'office, [**elle**] a saisi la chambre de recours.

Les solutions possibles que nous avons évoquées pour l'exemple (16) auraient permis d'éviter cette erreur, mais ce faisant, elles auraient masqué un défaut de la règle Z-PC.4.

Sous-estimation de la distance pronom-antécédent

Les erreurs au niveau des règles sur les zones d'antécédence consistent essentiellement à sous-estimer la possibilité qu'une expression pronominale renvoie à une expression de la phrase précédente.

Un pronom peut par ailleurs renvoyer à un antécédent se trouvant dans la deuxième phrase précédant celle du pronom. Seule la règle Z-PC.3 autorise une telle situation ; elle s'applique une fois dans le corpus d'étude et ne s'applique pas dans le corpus d'évaluation). On relève deux erreurs dues à cette limitation dans le corpus d'étude, aucune dans le corpus d'évaluation [17].

[17] Les deux extraits du corpus d'étude sont les suivants :

(18) + Le [**gouvernement**] attend cette semaine le dépôt des offres pour la reprise du GAN. + Et va ouvrir les chambres d'information sur la Marseillaise de Crédit. + Pour le Crédit Foncier, [**il**] commence l'analyse détaillée des propositions.

(19) Quatre [**administrateurs**] judiciaires ont été radiés. + La Commission nationale de discipline a rendu sa décision lundi dernier. + [**Ils**] ont refusé d'acquitter la cotisation exceptionnelle exigée par la profession pour combler le trou de l'étude Sauvan-Goulletquer.

Hormis les deux erreurs que l'on vient d'évoquer, où l'antécédent du pronom se trouve dans la deuxième phrase précédente, toutes les erreurs sur le corpus d'étude mettent en jeu un antécédent qui se trouve dans la phrase précédant celle de l'expression pronominale. La même faiblesse se retrouve pour 17 des 19 erreurs sur corpus d'évaluation [18].

Parmi ces erreurs, relevons deux cas où un pronom est identifié à tort comme cataphorique ; dans la phrase :

(20) Accusée, la [**Commission**] bancaire est vouée au silence en vertu du secret professionnel. Mais dans [**son**] entourage, on s'insurge.

son est analysé comme coréférent avec *on*, et dans l'exemple :

(21) A cet été, « [**elle**] redéfinira les politiques de maîtrise des risques, de distribution du crédit et de recouvrement des créances ». Une fois [**sa**] mission accomplie, un nouveau conseil d'administration sera élu.

sa est analysé comme coréférent avec *un nouveau conseil d'administration*.

À ces deux erreurs s'ajoute l'erreur suivante :

(22) Il est vrai que, indépendamment de la volonté du [**géant**] américain de se développer en France où [**il**] a déjà acquis le Crédit de l'Est, la Sovac, Locafrance et Ista, le Crédit Foncier ne constituait pas réellement un enjeu stratégique pour lui.

où le pronom *il* est analysé comme coréférent avec *le Crédit Foncier*.

Nous n'avons pas observé d'éléments qui caractérisent les quinze erreurs restantes, si ce n'est les quelques pistes suivantes.

De manière générale, on pourrait dire qu'une expression pronominale qui précède le verbe de la principale sans en dépendre directement peut avoir sa source dans la même phrase ou dans la phrase précédente, cela s'ajoutant à un éventuel cas de cataphore. Outre les exemples (20) et (21), ainsi que l'exemple (14), une telle règle serait susceptible de rendre compte des deux cas suivants :

(23) + La bonne tenue de l'activité, notamment de crédit, et la vigueur des marchés laissent augurer une nouvelle année de croissance des bénéfices pour les [**banques**] françaises. + Mais la situation en Asie, qui [**leur**] a déjà coûté cher en 1997, pourrait amoindrir leurs performances.

(24) Aux Pays-Bas, [**Hans Bartelds**], coprésident de Fortis, a fustigé la décision d'ABN-Amro jugée d'autant plus « inamicale », qu'elle tord le cou à la

Dans les deux cas, le contexte est celui du titre de l'article et/ou du chapeau. Dans (18), les deux premières phrases, distinguées typographiquement, sont en fait coordonnes. Dans (19), il est possible que la reprise constituée par le syntagme *sa décision* (qu'on peut expliciter comme sa décision [de radier les quatre administrateurs]) ait une influence dans l'interprétation du pronom *Ils*.

[18] Les erreurs déjà évoquées dans les sous-sections précédentes font partie de cet ensemble. Les deux erreurs restantes sont celles des exemples (11) et (22)

tradition consensuelle néerlandaise. A Bruxelles, Maurice Lippens, [**son**] homologue belge, manifestement confiant, a affirmé qu'une « surenchère est possible ».

On relève par ailleurs quelques cas où le verbe principal est impersonnel ou sans sujet :

(25) Dubitatif sur cette opinion, l'avocat Dominique Lefort estime que l'[**assistant**] spécialisé jouera un rôle proche de l'expert ou de l'officier de police judiciaire. Il faudra être « vigilant » sur les opinions qu'[**il**] pourra émettre, susceptibles selon lui d'« influencer le juge d'instruction ».

(26) Enfin, la [**banque**] discute actuellement avec le gouvernement d'un plan de recapitalisation de l'ordre de 2 000 milliards de wons dans les deux prochaines années. Reste à savoir si ces annonces spectaculaires [**lui**] permettront de redorer son blason international.

(27) L'un des « [**Big Six**] » vient de se voir refuser une offre de reprise d'un très grand cabinet d'avocats allemand. Il s'agit d'essayer via des acquisitions d'acheter une qualité de prestation qui [**leur**] échappe encore.

Notre définition des verbes « satellites » d'un autre verbe (voir chapitre 10 page 310) met précisément en jeu les verbes à sujet impersonnel. Les verbes dont dépendent les pronoms en question dans ces exemples ne sont pas satellites des verbes à sujet impersonnel parce qu'ils sont dans une proposition relative ou interrogative indirecte. Un élargissement de la notion de verbe satellite par l'utilisation d'un rapport matrice-enchâssée plutôt que par un rapport de complémentation entre les deux verbes serait susceptible de rendre compte de ces cas.

Ces développements, à supposer qu'ils soient pertinents, ne permettraient pas de résoudre correctement tous les cas dans notre corpus où, pour une expression pronominale, nos règles sur les zones d'antécédence ne spécifient pas un antécédent possible correct. Un travail reste donc à faire sur ce point.

12.3.3 Contraintes

Les contraintes (voir la section 10.3) se répartissent en contraintes d'accord, contraintes relationnelles et contraintes sur les insertions.

Les contraintes sont la cause de trois erreurs dans le corpus d'étude et de 7 erreurs dans le corpus d'évaluation (voir le tableau 12.8 page 386). Aucune de ces erreurs n'est dues aux contraintes relationnelles [19], quatre sont dues aux contraintes d'accord en genre, quatre aux contraintes d'accord en nombre et deux aux contraintes sur les insertions.

[19] Si on excepte les erreurs dues à un défaut dans la structure spécifiée par l'analyseur syntaxique.

Avant d'en venir à l'analyse de ces différentes erreurs dans les trois sous-sections suivantes, nous présentons la manière dont est calculé le taux de succès des contraintes.

TAUX DE SUCCÈS. Le taux de succès des contraintes S_c est déterminé comme le ratio suivant :

$$S_c = \frac{N - E_c}{N}$$

où N est le nombre total de reprises visées dans le corpus et E_c le nombre d'erreurs dues aux contraintes. Une contrainte commet une erreur si pour une expression pronominale e, elle élimine tous les antécédents possibles pour e, ces antécédents étant déterminés :

- par les règles sur les expressions dénotantes et les zones d'antécédence si celles-ci identifient au moins un antécédent correct pour e,
- en faisant abstraction des erreurs dues aux règles ou à l'analyseur pour les cas où il n'y a pas d'antécédent correct en sortie des règles [20].

Le taux de succès sur le corpus d'étude est de 99,23 % (385/388). Il est de 98,06 % sur le corpus d'évaluation (353/360). Comme pour les règles sur les zones d'antécédence, le taux d'erreur a plus que doublé (de 0,77 à 1,94).

Accord en genre

L'erreur d'accord en genre dans le corpus d'étude n'en est pas vraiment une. Dans le texte suivant, les expressions *le cabinet lyonnais Michaud*, *cette structure* et *il* constituent une chaîne de coréférence. Les règles sur les zones d'antécédence sélectionne *cette structure* comme antécédent possible pour *il*, ce qui est correct au regard de notre critère d'évaluation, mais cet antécédent est éliminé par les contraintes d'accord en genre. Le locuteur a probablement employé le pronom *il* (et utilisé la forme masculine *Fondé*) en ayant en tête la description *cabinet d'avocats*.

(28) Le cabinet d'avocat Mazars & Associés s'est rapproché du cabinet lyonnais Michaud. Fondé en 1981 par Pierre-Henry Michaud, cette [**structure**] compte 3 avocats. Spécialisé en droit des sociétés, [**il**] vient compléter l'activité de droit social de Mazars & Associés à Lyon.

Les erreurs d'accord en genre dans le corpus d'évaluation concernent toutes des noms propres. Dans la phrase suivante, le nom propre *Epicea* est donné par l'analyseur comme masculin. Le système retient comme antécédent le syntagme *la haute technologie*.

[20] L'analyse est ici faite dans le même esprit que celle des règles sur les zones d'antécédence (voir ci-dessus, page 390).

(29) Implantée à Paris, [**Epicea**] est spécialisée dans la haute technologie. [**Elle**] détient des participations dans 35 PME françaises, en particulier dans les secteurs des technologies de l'information et du biomédical.

Dans la phrase suivante, le syntagme *les AGF* est donné comme masculin par l'analyseur. La raison en est que un syntagme nominal pluriel dont le genre ne peut être déterminé par la forme du nom ou le déterminant est considéré par défaut comme masculin.

(30) + Les négociations vont reprendre de plus belle après la décision de Bruxelles d'exiger des [**AGF**] qu'[**elles**] cèdent leur participation de 24,8 % dans l'assureur crédit.

La règle voulant que le masculin « l'emporte » sur le féminin, cette préférence pour le masculin fait sens, ou moins statistiquement, pour les noms communs, comme, par exemple, *actionnaires* dans la phrase suivante, mais pas pour les noms propres.

(31) C'est d'ailleurs ce que les actionnaires exigent de leurs entreprises : qu'elles se donnent les moyens de disposer des fonds nécessaires pour satisfaire leurs besoins d'investissement et surmonter les crises.

La phrase de l'exemple (30) apparaît dans le titre d'un article et est reproduite presque à l'identique dans le corpus du texte, donnant lieu à notre troisième erreur d'accord en genre dans le corpus d'évaluation. Notons que dans les deux cas où le système échoue à trouver le lien entre le pronom *elles* et *les AGF*, le pronom est lui-même repris par un possessif, ce qui entraînera deux erreurs de chaînage en sortie finale du système.

Accord en nombre

Toutes les erreurs d'accord en nombre mettent en jeu un antécédent qui est un syntagme nominal singulier dénotant un ensemble de personnes au sens large [21].

(32) La ministre de la Justice Elisabeth Guigou a rappelé hier devant la commission des Lois qu'un décret circule au sein de la profession des mandataires et administrateurs liquidateurs des entreprises. « Ce décret vise à renforcer les contrôles sur cette [**profession**] et en particulier à [**les**] obliger à déposer leurs fonds auprès de la Caisse des dépôts », a-t-elle précisé.

(33) L'[**intersyndicale**] du GAN, qui a auditionné en début de semaine les quatre candidats à la reprise de [**leur**] groupe (Eureko, Groupama, AIG,

[21] L'exemple (33) contient deux erreurs d'accord en nombre ; le déterminant possessif *leur* ne peut être antécédent de *ils* car il figure dans une insertion, ce qui n'est pas faux dans la mesure où un antécédent correct (*l'intersyndicale*) figure dans l'ensemble des antécédents possibles de *ils*.

Swiss Life), estiment qu'[**ils**] ne seront pas en mesure de se prononcer valablement sur les candidatures lors du CCE de demain « compte tenu des fortes zones d'ombre qui demeurent sur certains projets ».

(34) Pour ce dernier, ce texte va notamment permettre à l'[**industrie**] pharmaceutique de trouver plus facilement des couvertures pour garantir [**leurs**] risques.

Le problème illustré par ces exemples est connu (voir, par exemple, [37, §629b] ou [9]). La solution passe probablement par le codage de l'information nécessaire sur un certain nombre d'entrées du lexique. Il s'agirait d'encoder le fait que tel ou tel nom, au singulier, peut dénoter un ensemble.

Contraintes sur les insertions

Deux erreurs sont dues aux contraintes sur les insertions. Dans l'exemple suivant, le syntagme *Gérard Mestrallet* figure dans l'insertion *menée par Gérard Mestrallet* et est donc exclu par le système comme antécédent possible de *son*. Peut-être le fait que cette insertion se trouve enchâssée dans le syntagme où figure le possessif justifierait-il une exception à notre règle.

(35) L'opération était en effet considérée comme une étape décisive dans la stratégie, menée par [**Gérard Mestrallet**], de recentrage de [**son**] groupe, qui passe inévitablement par une redéfinition de ses structures en Belgique.

Pour le possessif *leurs* dans l'exemple suivant, la zone d'antécédence est spécifiée comme la phrase courante. Le pronom *leur* figure dans une insertion et est donc exclu. Le fait que ce dernier soit une reprise pourrait justifier une exception à notre règle. Notons aussi que l'erreur pourrait aussi être analysée comme une erreur sur la zone d'antécédence.

(36) + La bonne tenue de l'activité, notamment de crédit, et la vigueur des marchés laissent augurer une nouvelle année de croissance des bénéfices pour les banques françaises. + Mais la situation en Asie, qui [**leur**] a déjà coûté cher en 1997, pourrait amoindrir [**leurs**] performances.

Contribution des différentes contraintes

Si les contraintes sont susceptibles d'engendrer quelques erreurs, il est intéressant de mesurer leur contribution respective à la tâche de résolution des pronoms. On le fait ici en suivant la démarche suivante.

On a en sortie des règles sur les zones d'antécédence un ensemble de couples (e, R_e) où e est une reprise et R_e est l'ensemble des référents possibles identifiés par le système pour cette reprise [22]. On ne considère que les couples pour lesquels

[22] Voir la section « Antécédents et référents », page 378, pour la définition de l'ensemble R_e.

	étude			*évaluation*		
	N_e	N_R/N_e	%R	N_e	N_R/N_e	%R
sortie zones d'antécédence	374	4,101		329	4,106	
contraintes relationnelles	373	3,831	6,6	327	3,902	5
accord en nombre	373	2,975	27,5	327	2,969	27,7
accord en genre	373	3,6	12,2	326	3,613	12
contraintes sur les insertions	373	3,75	8,6	327	3,807	7,3
ensemble des contraintes	370	2,281	44,4	320	2,343	42,9

TAB. 12.11 – Contribution des différentes contraintes.

il existe dans l'ensemble R_e un référent correct pour la reprise e. Soit N_e le nombre total de ces reprises correctes et N_R le nombre de total de référents possibles spécifié par le système pour l'ensemble de ces reprises. La division N_R/N_e donne le nombre moyen de référents possibles par reprise.

Pour le corpus d'évaluation, on a en sortie des règles sur les zones d'antécédence les valeurs suivantes, toutes erreurs prises en compte (c'est-à-dire erreurs de l'analyseur syntaxique incluses) :

$N_e = 329$
$N_R = 1351$
$N_R/N_e = 4,106$

Soit C un ensemble de contraintes. L'application de ces contraintes sur la sortie des règles sur les zones d'antécédence va donner lieu à une nouvelle sortie à partir de laquelle seront calculées de nouvelles valeurs N_e, N_R et N_R/N_e. La contribution d'un ensemble de contraintes C est mesurée par comparaison de ces valeurs : idéalement, N_e doit rester constant et plus la réduction de N_R est importante, plus la contrainte est intéressante.

Les résultats obtenus sont donnés dans le tableau 12.11. La ligne « sortie zones d'antécédence » reprend les chiffres donnés ci-dessus. Les cinq lignes suivantes donnent les chiffres obtenus par application de l'ensemble de contraintes correspondant *à la sortie des règles sur les zones d'antécédence* [23]. Un référent pour une reprise donnée peut être exclu par plusieurs contraintes différentes, si bien que la dernière ligne (contribution de l'ensemble des contraintes) n'est pas le résultat de la somme des différents apports mesurés dans les quatre lignes précédentes (contribution d'un groupe de contraintes particulier).

Toutes les erreurs comptabilisées pour les contraintes relationnelles sont dues à un défaut dans l'analyse syntaxique. L'erreur sur le pronom *ils* évoquée dans la note 21 page 399 est comptabilisée ici comme due aux contraintes sur les insertions.

[23] On évalue ici les contraintes indépendamment de l'ordre dans lequel elles s'appliquent.

Si on considère la seule réduction du nombre de référents potentiels, sans prendre en compte les éventuelles erreurs, l'accord en nombre est la contrainte la plus efficace, avec une réduction de l'ensemble des référents d'environ 27 %, devant l'accord en genre, qui à lui seul réduit l'ensemble des référents d'environ 12 %. La raison en est que l'information sur le nombre est associée à toutes les expressions pronominales, ce qui n'est pas le cas de l'information sur le genre. Un déterminant possessif, par exemple, ne donne pas d'indication sur le genre de son antécédent.

On remarquera que la contribution des différentes contraintes est globalement la même sur les deux corpus.

12.4 Évaluation des préférences

La section 12.4.1 décrit les données qui sont prises en compte pour l'évaluation des préférences. Les deux sections suivantes donnent respectivement une évaluation globale de l'ensemble des préférences et une évaluation analytique de chacune des préférences.

12.4.1 Données

Soit p_i une préférence ou un ensemble de préférences. La préférence p_i s'applique sur une sortie intermédiaire du système que nous noterons SI_{i-1} et produit une nouvelle sortie intermédiaire SI_i. Pour évaluer les préférences, nous raisonnerons en termes de « référents possibles » pour les expressions pronominales visées et non en termes d'« antécédents possibles » (voir la section 12.1.6). On représente les données d'une sortie intermédiaire SI comme un ensemble de couples (e, R_e) ou e est une expression pronominale et R_e est l'ensemble des référents possibles pour l'expression e selon le système.

Reprises ambiguës. On appelle N_{amb} l'ensemble des « reprises ambiguës » dans une sortie intermédiaire SI_{i-1}, sur laquelle s'applique une préférence p_i. Une reprise est une « reprise ambiguë » s'il existe plusieurs référents possibles pour cette reprise et si le référent correct figure parmi ces référents [24] :

$$N_{amb} = \{\ e \mid (e, R_e) \in SI_{i-1},\ |R_e| > 1,\ (\exists\ r \in R_e,\ r \subset E_e)\}$$

Application significative d'une préférence. Étant donné une sortie intermédiaire SI_{i-1}, une préférence p_i n'est susceptible de produire un résultat que nous jugerons significatif que pour les expressions qui appartiennent à N_{amb}. Une préférence s'applique de manière significative pour une expression pronominale e

[24] Dans la formule qui suit, la suite $\exists\ r \in R_e,\ r \subset E_e$ spécifie le critère de correction (voir la section 12.1.6) ; l'ensemble E_e est l'ensemble de référence associé à e dans la clé (voir la section « Information spécifiée pour les différents types d'expressions » p. 372).

si elle élimine un référent possible pour e et s'il existait un référent correct pour e avant application de p_i. Autrement dit, seront jugés non significatifs les cas d'application d'une préférence p_i suivants :

- si p_i élimine un référent incorrect pour une expression e, mais il n'existait pas de référent correct pour e avant application de p_i (évaluer le choix fait par p_i sur un ensemble de possibles qui sont tous incorrects n'a pas de sens),
- si p_i élimine un antécédent correct pour une expression e, mais un autre antécédent correct figure toujours dans l'ensemble des antécédents possibles pour e après application de p_i (ce cas de figure est illustré par l'application de la préférence 1 sur l'exemple (5), décrit page 352).

À ces conditions s'ajoute la condition suivante. Une préférence pose toujours des conditions sur l'existence de deux relations `coref(`e_i`,`e_j`)` et `coref(`e_i`,`e_k`)` et conclut à l'effacement de l'une de ces deux relations, mettons `coref(`e_i`,`e_j`)`. Le résultat de l'application d'une préférence p_i qui n'élimine pas le référent correct pour une reprise e_i sera jugé significatif si un antécédent correct pour cette reprise e_i a les propriétés requises pour instancier la variable e_k (c'est-à-dire le deuxième argument de la relation `coref` retenue lors d'une application de la préférence) dans les diverses applications de la préférence p_i pour l'expression e_i. En pratique, cette condition n'est pertinente que pour trois préférences : les préférences 7, 8 et 9. L'exemple suivant illustrera le cas de figure visé ici. Pour le déterminant possessif *sa* dans la phrase suivante, la préférence 9 élimine *une décision* de l'ensemble des antécédents possibles :

(37) La [société]^i fait appel d'une [décision]^j de l'[Ohmi]^k qui a rejeté [sa]$_{i/*j/k}$ demande d'enregistrement de marque communautaire.

Cependant, cet antécédent est éliminé sur la base d'un choix entre *une décision* et *l'Ohmi* et seulement entre ces deux expressions (la préférence 9 exprime une préférence pour le syntagme complément plutôt que le syntagme complété). On considère l'application de cette préférence pour cet exemple non significative, parce qu'elle ne fait pas intervenir l'antécédent réel de l'expression pronominale. La situation est similaire à celle d'un choix entre des antécédents qui sont tous erronés.

On note REP_{p_i} l'ensemble des reprises pour lesquelles une préférence ou un ensemble de préférences p_i s'applique de manière significative.

Application correcte ou incorrecte d'une préférence. Pour une reprise e de REP_{p_i}, la préférence p_i s'applique de manière incorrecte si elle élimine le référent correct pour e de l'ensemble R_e. On distingue deux sources d'erreurs : les erreurs dues à la préférence p_i elle-même et les erreurs dues à un défaut dans l'analyse syntaxique. On note le nombre d'erreurs dues à la préférence E_{p_i} et le nombre d'erreurs dues à l'analyseur E_a. Pour l'évaluation de la préférence p_i, nous ne tiendrons pas compte des expressions de REP_{p_i} pour lesquelles p_i

s'est appliquée en raison d'une erreur de l'analyseur. On se donne donc la valeur $N_{p_i} = |REP_{p_i}| - E_a$, qui servira de dénominateur pour calculer le taux de succès de la préférence p_i.

TAUX DE SUCCÈS. Le taux de succès s_{p_i} d'une préférence ou d'un ensemble de préférences p_i correspond au rapport suivant :

$$s_{p_i} = \frac{N_{p_i} - E_{p_i}}{N_{p_i}}$$

c'est-à-dire le rapport du nombre de reprises pour lesquelles p_i s'est correctement appliquée sur le nombre total de reprises pour lesquelles p_i s'est appliquée, abstraction faite des cas où p_i s'est appliquée suite à une erreur de l'analyseur syntaxique.

DÉTERMINATION DES VALEURS POUR L'ÉVALUATION D'UNE PRÉFÉRENCE. Les différentes valeurs que nous avons présentées nous servirons à évaluer les préférences. Si pour les règles nous avons pu systématiquement faire abstraction des erreurs dues à l'analyseur et/ou des erreurs dues aux autres composants du système de résolution, il n'en est pas de même pour l'évaluation des préférences. Une préférence ou un ensemble de préférences p_i sera analysé sur l'ensemble des reprises pour lesquelles il existe, selon le système réel, une réponse correcte dans la sortie intermédiaire qui précède immédiatement l'application de la préférence p_i. On veut dire par là que si, pour une raison ou une autre, le système ne donne pas de réponse ou donne une réponse incorrecte pour une expression e avant l'application de p_i, le résultat qu'*aurait donné* l'application de p_i n'est pas évalué pour cette expression. Les chiffres donnés ci-après pour l'évaluation des préférences doivent donc être lus comme des indications calculées sur un échantillon du corpus, non comme des valeurs absolues sur l'ensemble du corpus.

12.4.2 Évaluation globale

On évalue d'abord les préférences de manière globale, c'est-à-dire que p_i est l'ensemble de toutes les préférences et N_{amb} est l'ensemble des reprises ambiguës en sortie des règles. Nous dirons des valeurs obtenues qu'elles caractérisent le « taux de succès global » (noté s_g) des préférences.

Pour le corpus d'étude, on a 209 expressions ambiguës en sortie des règles. Sur ces 209 expressions, l'application de l'ensemble des préférences produit 188 réponses correctes et 21 erreurs, dont 17 sont dues aux préférences elles-mêmes et 4 à l'analyseur. Le taux de succès global est donc, pour le corpus d'étude :

$$s_g = \frac{188}{209 - 4} = 91,7\ \%$$

Pour le corpus d'évaluation, on a 178 expressions ambiguës en sortie des règles. Sur ces 178 expressions, l'application de l'ensemble des préférences produit 146

réponses correctes et 32 erreurs, dont 23 sont dues aux préférences elles-mêmes et 9 à l'analyseur. Le taux de succès global est donc, pour le corpus d'étude :

$$s_g = \frac{146}{178 - 9} = 86,39\ \%$$

Comme pour les règles sur les zones d'antécédence et les contraintes, on constate une baisse sensible des résultats entre le corpus d'étude et le corpus d'évaluation.

12.4.3 Évaluation analytique

Après l'évaluation globale des préférences, nous évaluons ici tour à tour chacune des préférences.

Les préférences sont ordonnées, si bien que chaque préférence p_i doit être évaluée relativement à l'ensemble des reprises ambiguës après application de la préférence précédente p_{i-1}.

Les résultats pour chacune des préférences, évaluées compte tenu de l'ordre dans lequel elles s'appliquent, sont donnés dans le tableau 12.12 pour le corpus d'étude et pour le corpus d'évaluation. Chaque ligne identifie une préférence ou un ensemble de préférence p_i [25]. La colonne N_{amb} donne le nombre de reprises ambiguës avant application de la préférence en question. La colonne N_{p_i} donne le nombre d'expressions pronominales pour lesquelles la préférence p_i s'applique, abstraction faite des cas où elle s'applique suite à une erreur de l'analyseur. La colonne E_{p_i} donne le nombre d'erreurs commises pour les expressions de N_{p_i} par faute de la préférence. Enfin la colonne s donne le taux de succès de la préférence p_i (voir ci-dessus comment le taux de succès est calculé).

Premières préférences. Les cinq premières préférences semblent avoir une certaine validité, qui doit cependant être relativisée compte tenu du fait qu'elles ne s'appliquent que pour un faible nombre d'expressions pronominales (un cas extrême est celui de la préférence pour le parallélisme des fonctions (5), qui ne s'applique pas sur le corpus d'évaluation). On notera que l'erreur attribuée à la préférence 4 sur le corpus d'étude pourrait être attribuée à l'analyseur. Dans la phrase

(38) Une semaine après s'être divisés sur le [projet]^i de fusion avec [Fortis]^j, les actionnaires de la [Générale]^k de Banque [l']$_{*i/j/k}$ ont approuvé hier à l'unanimité.

le syntagme *sur le projet* est relié au verbe *s'être divisés* par une relation `vmod` alors qu'une relation `varg` serait ici peut-être plus appropriée (voir la description

[25] Les chiffres sont ceux qui référencent les préférences dans le chapitre 11.

	corpus d'étude				*corpus d'évaluation*			
Préf.	N_{amb}	N_{p_i}	E_{p_i}	s	N_{amb}	N_{p_i}	E_{p_i}	s
1	209	5	0	100	178	4	1	75
2	207	3	0	100	177	2	0	100
3	204	17	0	100	175	12	0	100
4	199	27	1	96,3	171	14	0	100
5	192	3	0	100	169	0	0	–
6	189	12	1	91,7	169	16	1	93,8
7	183	39	1	97,4	159	23	0	100
8	152	43	3	93	140	48	2	95,8
9	129	18	1	94,4	118	12	1	91,7
10	121	65	2	96,9	110	50	2	96
11	73	13	1	92,3	77	19	5	73,7
12	61	11	0	100	58	4	2	50
13	50	12	2	83,3	54	12	1	91,7
14	39	32	3	90,6	41	33	6	84,8
15	10	9	2	77,8	9	8	2	75

TAB. 12.12 – Évaluation des préférences.

de ces relations page 256). Quoiqu'il en soit, le syntagme *sur le projet* n'est pas un complément de lieu, expressions *a priori* visées par la préférence en question [26].

Par ailleurs, l'erreur due à la première préférence sur le corpus d'évaluation pourrait être évitée par un changement dans l'ordre des préférences. Dans la phrase suivante, avant application de la préférence 1, le pronom *elle* a deux antécédents possibles : *ses* et *Suisse*.

(39) La banque japonaise Sumitomo a annoncé son intention de fermer sa filiale suisse Sumitomo Bank (Schweiz) dans le cadre de la restructuration de ses activités à l'international. A l'avenir, [ses]^i activités en [Suisse]^j, dont notamment l'émission et la souscription de valeurs mobilières, seront assurées par la Banque du Gothard, dont [elle]$_{*i/j}$ détient 55 %.

Le déterminant possessif *ses* est éliminé au profit de *Suisse*, mais si la préférence 1 était appliquée après la préférence 4, alors *Suisse* serait éliminé avant le possessif et la réponse du système serait alors correcte.

Cette erreur invite donc à redéfinir l'ordre des premières préférences de 1-2-3-4 à 2-3-4-1.

[26] Nous attribuons cette erreur à la préférence 4 parce que, d'une part, l'opérationalité de la distinction entre `vmod` et `varg` n'est selon nous pas établie, d'autre part, si la préférence 4 vise des compléments de lieu, elle le fait de façon approximative et c'est cette approximation qui est ici en cause.

Sur la préférence 6. Les conditions posées par la préférence 6 sont de manière générale posées par ailleurs par d'autres préférences. Nous ne sommes pas en mesure de déterminer si tout ce qu'exprime cette préférence est exprimé par les préférences qui suivent, mais le fait est que la suppression de la préférence 6 ne change le résultat final sur aucun des deux corpus. En pratique, les cas traités dans les deux corpus par la préférence 6 sont couverts par les préférences 7, 8, 10, 11 et 13.

Les deux phrases pour lesquelles cette préférence cause une erreur sont intéressantes dans la mesure où elles contredisent plusieurs préférences générales. Dans l'exemple (40), l'antécédent du pronom il_2 (en l'occurrence *le dossier*) n'est ni le plus proche, ni une reprise, ni un syntagme qui dénote une personne. Les préférences aboutiront à tort à la conclusion que il_2 est coréférent avec *le gouvernement* et il_1.

(40) Reste enfin le dossier de la rémunération des dépôts que le gouvernement regarde de très près, même s'il$_1$ souligne qu'il$_2$ est plus du ressort de la place que du sien.

De manière similaire, en (41), l'antécédent du déterminant possessif son_1 (en l'occurrence *un client*) n'est ni sujet, ni une reprise, ni le plus proche antécédent possible pour cette expression. Les préférences aboutiront à la conclusion que son_1 est coréférent avec *un auditeur légal*. Le second possessif est quant à lui mal interprété par la préférence 11, qui sélectionne *le cabinet* comme antécédent pour cette expression.

(41) A titre d'anecdote, un auditeur légal a déclaré récemment à un client qu'il ne certifierait les comptes de son$_1$ groupe que si le cabinet d'avocats membre de son$_2$ réseau était retenu sur un dossier d'acquisition en cours...

Préférences les plus significatives. Nous considérons les préférences 7 (préférence pour le sujet dans les reprises inter-phrases), 8 (préférence pour le sujet et la proximité dans les reprises intra-phrases) et 10 (préférence pour une personne et cohésion dans les reprises par pronom datif ou déterminant possessif) comme les plus significatives, dans le sens où elles s'appliquent sur un nombre relativement important d'expressions avec une validité assez forte.

La préférence 7 a une forte validité sur les deux corpus (97,4 % et 100 %). On notera que cette validité est supérieure à celle de la préférence 8 ce qui confirme l'idée (exposée page 359) que la préférence pour le sujet est plus nette dans les reprises inter-phrases que dans les reprises internes à la phrase (d'autant plus que la préférence 8 ne sélectionne pas seulement des expressions sujet).

L'erreur due à la préférence 7 sur le corpus d'étude est la suivante. Dans le texte suivant, le syntagme *le Crédit Lyonnais* est retenu comme antécédent de *Il* :

(42) Ce qui conduit à penser que le [Crédit]^i Lyonnais va devoir émettre un [montant]^j significatif de nouveaux titres pour financer ce [rachat]^k. [Il]$_{i/*j/*k}$ pourrait se situer autour de 25 milliards de francs.

On notera que si cet exemple contredit la préférence 7, il contredit aussi toute préférence pour un syntagme dénotant une personne. Pour résoudre cet exemple, il faudrait utiliser l'information selon laquelle le prédicat exprimé par le syntagme verbal se dit plutôt d'un montant que d'une société.

On notera par ailleurs que dans l'exemple (42), le syntagme *le Crédit Lyonnais* est sujet mais n'est pas sujet d'une proposition principale. Sur l'ensemble des reprises résolues par la préférence 7, 57 le sont par rapport à une expression sujet de la proposition principale. Ces 57 expressions étant toutes interprétées correctement, on a peut-être là une condition suffisante (si une expression pronominale e_i peut renvoyer [27] à un syntagme nominal e_j sujet de la proposition principale de la phrase précédente, alors e_i est de préférence coréférente avec e_j) permettant de rendre compte de façon très fiable de l'interprétation de certaines expressions.

La forte validité de la préférence 7 illustre bien selon nous l'intérêt des préférences ordonnées par rapport à des préférences pondérées (voir la section 11.1.1). C'est parce que nos préférences expriment un choix binaire sur des paires d'antécédents possibles pour une expression pronominale, que nous pouvons aujourd'hui faire l'hypothèse d'une condition suffisante pour déterminer correctement et de manière fiable l'interprétation de certaines expressions pronominales dans certains contextes. L'usage de préférences pondérées aurait sans doute moins bien permis de faire ressortir la régularité observée ici.

Pour terminer cette discussion de la préférence 7, signalons pour l'anecdote que nous avons présenté à huit personnes le texte incomplet :

(43) Ce qui conduit à penser que le Crédit Lyonnais va devoir émettre un montant significatif de nouveaux titres pour financer ce rachat. Il pourrait...

en leur demandant de choisir entre les trois antécédents possibles. Sept personnes ont choisi *le Crédit Lyonnais*, une a choisi *un montant*, ce qui va dans le sens de notre préférence.

Parmi les erreurs dues à la préférence 8, notons les deux suivantes. Dans l'exemple (44), le syntagme *le Crédit Lyonnais* est retenu comme antécédent de *son* (dans *son coût*). Notons que l'antécédent correct (*du prêt à l'établissement public de financement et de restructuration*) serait aussi éliminé par la préférence pour un antécédent dénotant une personne. Outre le fait que le nom *coût* sélectionne sans doute de préférence un syntagme dénotant un prêt plutôt qu'une société, on remarquera que l'apposition qui modifie le syntagme antécédent indique une certaine focalisation du discours sur le prêt en question.

(44) Il s'agit d'une part du rachat de la clause de retour à meilleure fortune, une « créance » instaurée par l'Etat en 1995, et d'autre part de la neutralisation du prêt à l'établissement public de financement et de restructu-

[27] Le terme « peut renvoyer » s'entend compte tenu de toutes les conditions posées dans notre système préalablement à l'application de la préférence 7.

ration, prêt actuellement rémunéré à seulement 85 % du taux du marché monétaire au Crédit Lyonnais qui ne parvient pas à couvrir son coût de refinancement.

La préférence 8 s'applique également dans les cas de cataphore potentielle. Dans l'exemple (45), l'expression *chacune* est éliminée parce elle précède un sujet avec lequel *ses* peut être coréférent. Une erreur due à la préférence 8 est évitée sur chacun des corpus si on exige que l'antécédent retenu précède le déterminant possessif.

(45) Autant dire qu'en dépit des assurances données par chacune^{i} sur la qualité de ses$_{*i/j/k}$ contreparties, le sort^{j} de ce pays^{k} constitue un enjeu de tout premier plan pour les banques françaises.

Sur les deux corpus, la préférence qui s'applique le plus souvent est la préférence 10 (préférence pour un syntagme dénotant une personne ou un pronom dans les reprises par pronom datif ou déterminant possessif), avec une fiabilité constante d'un corpus à l'autre. Nous avons vu que l'ordre des préférences 9 et 10 pouvaient être inversé sans entraîner de changement dans le résultat final (voir page 345). Appliquée sur la sortie de la préférence 8, la préférence 10 montre une validité supérieure à celle qui est mesurée dans le tableau 12.12 : 97,2 % sur le corpus d'étude et 96,4 % sur le corpus d'évaluation. Inversement, la préférence 9, appliquée à la sortie de la préférence 10, se révèle alors moins pertinente : elle ne s'applique plus que sur 13 reprises du corpus d'étude, avec un succès de 92,3 %, et sur 7 reprises du corpus d'évaluation, avec un succès de 85,7 %.

Autres préférences. Les taux de succès obtenus pour les cinq dernières préférences sont globalement inférieurs à ceux obtenus pour les autres préférences. De plus, ces préférences s'appliquent sur un ensemble assez réduit d'expressions ambiguës en entrée, ce qui rend leur évaluation moins pertinente.

En complément des chiffres présentés dans le tableau 12.12, nous présentons ici une manière complémentaire d'évaluer les dernières préférences du système (c'est-à-dire la préférence 9 et les préférences 11 à 14), en regardant dans quelle mesure le résultat final change si on supprime telle ou telle préférence.

On considère la sortie S_{p10} du système après application des préférences 1 à 10 moins la préférence 6, dont on a vu qu'elle pouvait être supprimée, et la préférence 9, dont a vu qu'elle pouvait s'appliquer après la préférence 10 sans changer les résultats. Dans chacun des deux corpus, la sortie S_{p10} contient 84 reprises ambiguës. Dans le corpus d'étude, on a en moyenne 2,417 référents possibles pour ces 84 reprises ; dans le corpus d'évaluation, le nombre moyen de référents possibles est de 2,238. À supposer qu'on sélectionne au hasard un référent pour chacune de ces reprises, on devrait obtenir un taux de succès proche des valeurs suivantes :

- 41,4 % pour le corpus d'étude,
- 44,7 % pour le corpus d'évaluation.

L'application de la suite de préférences 9-11-12-13-14-15 sur ces reprises ambiguës produit les résultats suivants [28] :

- 73 reprises correctement interprétées dans le corpus d'étude, soit un taux de succès de 86,9 % ;
- 63 reprises correctement interprétées dans le corpus d'évaluation, soit un taux de succès de 75 %.

Ces chiffres sont nettement supérieurs à ceux qui seraient obtenus par hasard, ce qui indique une certaine pertinence des préférences en question. La différence de plus de dix points entre les résultats obtenus sur le corpus d'étude et ceux qui sont obtenus sur le corpus d'évaluation tend à indiquer un certain degré de spécialisation des préférences par rapport au corpus d'étude.

On notera que la seule application des préférences finales (préférence 15) produirait les résultats suivants :

- 34 reprises correctement interprétées dans le corpus d'étude, soit un taux de succès de 40,5 % ;
- 33 reprises correctement interprétées dans le corpus d'évaluation, soit un taux de succès de 39,3 %.

Le chiffre est proche du résultat obtenu par hasard pour le corpus d'étude et nettement inférieur au hasard pour le corpus d'évaluation. Appliquée sur la sortie S_{p10}, la préférence 15 n'est donc pas pertinente ; en revanche les chiffres du tableau 12.12 montrent qu'elle l'est lorsqu'appliquée sur la sortie de la préférence 14.

En complément des mesures du taux de succès des préférences présentées dans le tableau 12.12, nous pouvons mesurer l'apport des différentes préférences au résultat final (RF), obtenu avec la suite de préférences 9-11-12-13-14-15, en supprimant tour à tour chacune [29] des préférences p_i et en comparant le résultat R_{-p_i} obtenu avec RF. Le tableau 12.13 donne les chiffres obtenus selon cette procédure. Les colonnes c et s donnent respectivement le nombre de reprises correctement résolues et le taux de succès correspondant, calculé sur l'ensemble des reprises ambiguës dans la sortie S_{p10} (rappel : dans chacun des deux corpus, on compte 84 reprises ambiguës). La colonne d (« différence ») donne la différence entre le nombre de reprises correctement résolues en sortie finale et le nombre de reprises résolues correctement sans la préférence i référencée dans la ligne R_{-i}. La ligne RF donne le résultat final obtenu avec l'application de la suite de préférences 9-11-12-13-14-15. Les lignes R_{-9} à R_{-14} donnent le résultat obtenu par application de la même suite moins la préférence i référencée dans R_{-i}. Les cinq dernières lignes donnent les résultats obtenus par suppression de certaines parties de la préférence 14, qui se trouve être la plus importante pour les deux

[28] Ces résultats incluent les erreurs éventuellement dues à l'analyseur.

[29] À l'exception de la préférence 15, qui est celle qui permet d'avoir une réponse unique pour chaque expression pronominale.

	corpus d'étude			*corpus d'évaluation*		
	c	d	s	c	d	s
RF	73	–	86,9	63	–	75
R_{-9}	70	−3	83,3	63	0	75
R_{-11}	70	−3	83,3	62	−1	73,8
R_{-12}	72	−1	85,7	63	0	75
R_{-13}	63	−10	75	59	−4	70,2
R_{-14}	56	−17	66,7	44	−19	52,4
$R_{-14\text{bcd}}$	64	−9	76,2	55	−8	65,5
$R_{-14\text{b}}$	67	−6	79,8	65	+2	77,4
$R_{-14\text{c}}$	73	0	86,9	64	+1	76,2
$R_{-14\text{d}}$	70	−3	83,3	61	−2	72,6
$R_{-14\text{bc}}$	67	−6	79,8	66	+3	78,6

TAB. 12.13 – Résultats après suppression de certaines préférences.

corpus.

Les résultats présentés dans les lignes R_{-9} à R_{-14} indiquent que toutes les préférences sont pertinentes pour le résultat final, dans le sens où la suppression d'une quelconque des préférences dégrade le résultat final sur l'un ou l'autre des deux corpus, sinon sur les deux. Les préférences 9, 11 et 12 ont une pertinence assez faible (en ce qui concerne 9 et 12, leur pertinence est nulle sur le corpus d'évaluation), sans doute pour deux raisons : d'une part, elles s'appliquent sur peu d'expressions (voir les chiffres des colonnes N_{p_i} dans le tableau 12.12), d'autre part, les préférences 11 et 12 expriment une préférence pour le sujet qui est également exprimée par la préférence 14 ainsi que, dans le cas de la préférence 11, une préférence pour la proximité qui est exprimée par la préférence 15.

Compte tenu du relativement faible nombre de cas où elle s'applique (voir le tableau 12.12), la préférence 13 (préférence pour la cohésion) s'avère assez pertinente. La préférence 14 est quant à elle la plus importante pour le résultat final, en particulier parce qu'elle est celle qui s'applique sur le plus d'expressions.

L'analyse des différentes préférences constituant le groupe de préférences 14 (cinq dernières lignes du tableau) révèle que les préférences 14b et 14c ont un effet négatif sur le corpus d'évaluation : la suppression de ces deux préférences conduit à une amélioration du résultat final (le nombre de reprises correctement interprétées augmente de 3). On notera cependant que dans le même temps, la suppression des préférences 14b et 14c (ou simplement 14b, puisque l'influence de 14c est nulle) dégrade les résultats sur le corpus d'évaluation. On a là un indicateur de la spécialisation des préférences par rapport au corpus d'étude.

CONCLUSIONS SUR LES PRÉFÉRENCES. Il ne nous semble pas nécessaire d'entrer plus avant dans l'analyse des résultats produits par les préférences que nous avons

définies. Nous avons dit (au début de la section 11.1) que les préférences avaient pour nous un caractère exploratoire. Les conclusions que nous pouvons tirer de cette exploration sont les suivantes.

De manière générale, les préférences ont une certaine validité statistique à la fois sur le corpus d'étude et sur le corpus d'évaluation.

Dans le même temps, cette validité est clairement seulement statistique : pour la majorité des préférences, le corpus lui-même contient un ou plusieurs contre-exemples. Nous en avons vu quelques-uns plus haut, en particulier les exemples (40) et (41), qui contredisent à eux seuls plusieurs préférences. Les préférences sont donc inadéquates pour décrire précisément l'interprétation des pronoms. Une des raisons de cette inadéquation vient de ce que l'interprétation des pronoms met probablement en jeu certaines sources d'information dont nous n'avons pas pu tenir compte. Parmi ces absences, mentionnons les restrictions de sélection (voir la section 6.2 et la discussion des exemples (42) et (44)) ou une information de nature pragmatique [30] (voir la section 6.3).

En l'absence de ces deux sources d'informations, ainsi que, peut-être, d'autres sources d'informations à découvrir, il semble qu'il est impossible de parvenir à la définition d'un système d'interprétation automatique des expressions pronominales qui soit réellement fiable. Cela étant, les préférences que nous avons définies nous semblent constituer une approximation intéressante pour un système susceptible d'être utilisé dans une application effective telle que la recherche d'information dans des bases de données textuelles.

12.5 Pertinence de l'évaluation

Le système que nous avons décrit a été défini à partir d'observations faites sur un corpus d'un certain type (articles du domaine de la finance) et évalué sur le même type de corpus. Cela étant plusieurs questions restent ouvertes :

1. Quels seraient les résultats sur un corpus du même domaine mais de source différente (par exemple, des articles du Monde traitant d'économie) ?
2. Quels seraient les résultats sur un corpus traitant d'un domaine différent, mais dans un style *a priori* similaire (par exemple, des articles de presse traitant d'un sujet autre que l'économie) ?
3. Quels seraient les résultats sur un corpus différent à la fois quand au style et quand au sujet (par exemple, un manuel technique, un roman) ?

L'annotation de corpus est un travail extrêmement coûteux en temps et il ne nous a pas été possible de mener dans le cadre de la thèse des expériences qui apportent une réponse à ces interrogations [31].

[30] L'interprétation de l'exemple (41) met peut-être en jeu ce type d'information.

[31] Le corpus annoté dans le projet décrit dans [90] pourra servir de base à une telle étude, mais il nécessite d'être adapté à la tâche que nous nous sommes fixée (il faut identifier les reprises

	A	B
Mitkov	48,57	71,76
Baldwin	42,85	67,05
Kennedy & Boguraev	55,71	74,11

TAB. 12.14 – Évaluation de trois systèmes sur deux textes différents.

La question de la pertinence de l'évaluation soulevée ici est d'autant plus cruciale si on considère quelques résultats obtenus avec différents systèmes de résolution sur différents corpus.

C. Barbu et R. Mitkov [8] ont développé un « banc d'évaluation pour la résolution d'anaphore qui permet l'incorporation de différents algorithmes et leur comparaison en utilisant les mêmes outils de pré-traitement et sur les mêmes données ». Dans ce contexte, trois systèmes ont été réimplantés et évalués sur la même analyse syntaxique en entrée : celui de Mitkov lui-même, celui de Baldwin et celui de Kennedy et Boguraev (voir les sections consacrées à ces systèmes dans le chapitre 6). Le corpus est constitué de quatre textes techniques du domaine de l'informatique. Au total, il contient 422 pronoms dont 362 sont anaphoriques.

Un premier point à noter est que, de manière générale, les résultats obtenus pour les trois systèmes sont sensiblement moins bons que ceux qui ont été décrits par leurs auteurs respectifs. Il est difficile de cerner la cause réelle de cette dégradation, qui peut en particulier venir de la mauvaise qualité des outils de pré-traitement (la sortie de ces outils ne subit aucune correction manuelle), mais elle n'en soulève pas moins une interrogation sur la variation des résultats d'un type de corpus à l'autre.

Deuxième point à noter, les résultats décrits dans [8] montrent des écarts assez nets d'un texte à l'autre pour un même système. Ainsi pour deux textes A et B contenant un nombre de pronoms comparable (92 et 97 pronoms, respectivement), les résultats sont ceux qui sont présentés dans le tableau 12.14 [32]. Pour les trois systèmes, l'écart entre les résultats obtenus sur le texte B et ceux qui sont obtenus sur le texte A est très important (de l'ordre d'une vingtaine de points). Là aussi, ces chiffres suggèrent que les résultats obtenus par un même système sur des corpus différents seront susceptibles de varier sensiblement.

Un de nos objectifs sera de tester notre propre système sur un ensemble de textes d'origines, de styles et de domaines variés, avec dans l'idée d'isoler les hypothèses qui sont valides à un niveau général (c'est-à-dire valides sur un grand nombre de textes différents) et celles qui sont plus spécifiques à un corpus.

pertinentes pour notre système) et au format de sortie de notre système (il faut identifier les noyau des syntagmes nominaux).

[32] La manière dont ces chiffres sont calculés est présentée dans [8]. Ce qui nous intéresse ici est surtout l'écart entre les chiffres pour un même système sur deux textes différents.

12.6 Conclusion et perspectives

Au terme de cette seconde et dernière partie de la thèse, nous faisons le bilan de notre travail et envisageons quelques perspectives pour l'avenir.

12.6.1 Apports de notre travail

Le système d'hypothèses que nous avons présenté et évalué ne permet pas de rendre compte complètement de l'interprétation des expressions pronominales, mais il permet d'en isoler certains aspects. En implantant notre système d'hypothèses, nous avons explicité des notions dont on ne trouve dans la littérature que des formulations intuitives. Par exemple, on sait qu'une expression pronominale e_i s'interprète comme dénotant un être désigné par ailleurs par une expression e_j apparaissant dans le contexte proche de l'expression pronominale et que les expressions e_i et e_j s'accordent en règle générale en genre et en nombre. Notre système d'interprétation automatique des expressions pronominales explicite ce que signifie les expressions « contexte proche » et « s'accorder en genre et en nombre ».

Dans la mesure où bon nombre des hypothèses implantées dans notre système traduisent des notions intuitives et connues, le lecteur aura peut-être l'impression de ne pas avoir appris grand-chose sur l'interprétation des expressions pronominales. Le fait de traduire ces intuitions et connaissance dans un système explicite, structuré et formalisé nous semble cependant un apport important de notre travail.

Pour illustrer la nécessité de l'explicitation de ce qu'on peut appeler les connaissances de base sur l'interprétation des pronoms (la majorité des connaissances exprimées par les règles dans notre système), on peut mettre en avant ce qui nous semble être un oubli dans l'exposé des règles de la théorie du centrage (voir la section 6.4.1). Les auteurs de la théorie du centrage se donnent quatre types de transitions entre énoncés et une hiérarchie de ces transitions censée refléter la cohérence du discours considéré. Les exemples (13) et (14) page 199 illustrent la cohérence relative de deux discours. Une question qu'on est en droit de se poser, au vu des exposés de la théorie du centrage dans [39] et [93], est celle de savoir si la hiérarchie des transitions, et donc les prédictions sur la cohérence relative d'un discours, valent dans tous les cas ou seulement lorsqu'il y a ambiguïté potentielle dans l'interprétation d'un pronom. Considérons le texte suivant :

(46) Le président de l'AFB, Michel Freyche, a indiqué hier que les négociations avec les syndicats sur une nouvelle convention collective reprendront début juin. Elles s'articuleront autour des quatre groupes de travail déjà définis. Ils devraient se retrouver « fin septembre-début octobre » pour une réunion de synthèse avant de reprendre les négociations par groupe.

Les transitions entre les énoncés de ce texte sont les mêmes que celles de l'exemple (14) p. 199, jugé par les auteurs de la théorie du centrage comme un discours relativement « coûteux » à comprendre. Une différence importante entre le texte de (14) p. 199 et le texte de (46) est cependant que les deux pronoms de (46) ne sont pas ambigus en sortie des règles de notre système. Quelles sont les prédictions de la théorie du centrage dans ce cas : le texte de (46) est-il aussi coûteux ou moins coûteux à comprendre que le texte de (14) p. 199 ? Notre hypothèse est que c'est la seconde de ces deux options qui est à retenir, mais les exposés de la théorie du centrage n'explicitent pas ce point.

Lorsqu'il s'agira d'explorer certains mécanismes en jeu dans l'interprétation des pronoms (par exemple, les restrictions de sélection ou le centrage), notre système pourra servir à isoler en corpus les cas pertinents pour cette exploration.

Au-delà de l'explicitation d'un certain nombre de connaissances intuitives sur l'interprétation des expressions pronominales, l'évaluation de notre système d'interprétation des expressions pronominales a mis à jour certaines connaissances que nous considérons comme moins évidentes.

Parmi celles-ci, l'idée, exprimée par les règles sur les expressions dénotantes, que l'antécédent d'une expression pronominale doit être un syntagme nominal accompagné d'un déterminant ou un nom propre méritait à nos yeux d'être exposée et testée. Nous admettrons volontiers que cette idée n'est pas une grande découverte, mais nous ne l'avons vue exposée nulle part de cette manière. Notons par ailleurs que le fait que deux contre-exemples aient été rencontrés dans notre corpus d'étude (voir p. 388) est une autre illustration de l'intérêt d'un système d'interprétation automatique effectif : il permet de faire apparaître des régularités et *a contrario* les irrégularités sur lesquelles il sera intéressant de se pencher dans l'avenir.

Parmi les contraintes, les contraintes sur les insertions (voir section 10.3.3) expriment des impossibilités de coréférence entre une expression pronominale et un antécédent potentiel qui n'ont a notre connaissance pas été exprimées et implantées comme nous l'avons fait. Ces contraintes ont une certaines validité, malgré deux contre-exemples qui seront à analyser, comme les contre-exemples aux règles sur les expressions dénotantes. On notera que nos contraintes sur les insertions constituent des contraintes différentes de celles qu'expriment les contraintes de liage ou de c-commande (c'est-à-dire que ces dernières n'incluent pas nos contraintes sur les insertions).

Par ailleurs, on pourra objecter que les contraintes sur les insertions relèvent d'un principe plus général formulé dans la théorie des veines (voir section 6.4.2). C'est sans doute le cas, mais nous avons montré que le principe général de la théorie des veines n'exprimait que l'hypothèse d'une corrélation entre structures rhétoriques et liens de coréférence, plutôt qu'une dépendance (à partir de laquelle on pourrait faire des prédictions) des liens de coréférence par rapport à des structures rhétoriques identifiables indépendamment des liens de coréférence (voir la

discussion p. 206). Nos contraintes sur les insertions sont donc beaucoup plus spécifiques que le principe formulé dans la théorie des veines, mais elles expriment des hypothèses effectivement testables parce que reposant sur une information obtenue indépendamment de l'hypothèse elle-même.

Parmi les préférences, nous avons vu que les premières préférences définies avaient une forte validité. Ces préférences n'expriment pas une connaissance particulièrement nouvelle, mais là encore il était utile d'expliciter cette connaissance. Parmi les préférences qui ont une certaine validité, la préférence 7 nous semble la plus intéressante, dans la mesure où elle exprime une connaissance moins évidente. Si cette connaissance n'est pas absolument nouvelle, puisque, comme nous l'avons dit, l'hypothèse en question a déjà été formulée par Baldwin [6], elle méritait (et mériterait encore) d'être confirmée.

Au-delà des connaissances qu'il explicite, notre système met en lumière certaines limites dans les connaissances actuelles sur l'interprétation des expressions pronominales. Ce point ne constitue pas le moindre des ses intérêts.

Enfin, pour terminer cet inventaire de ce que nous considérons être les apports de notre travail, nous insisterons sur le fait que notre système d'interprétation des expressions pronominales constitue un système effectivement utilisable et susceptible de produire des résultats intéressants pour une application de traitement automatique des langues telle que la recherche d'information dans des bases de données textuelles.

12.6.2 Perspectives

Le système d'interprétation des expressions pronominales que nous avons défini n'est bien évidemment pas parfait. Nous envisageons dans cette section les perspectives d'extension et d'amélioration pour l'avenir.

Affinage des hypothèses existantes

Nous avons dit plus haut, au terme de l'évaluation des préférences, qu'étant donné les sources d'information utilisées, il semblait impossible de parvenir à la définition d'un système vraiment fiable d'interprétation automatique des expressions pronominales. Ce constat nous invite à envisager une évolution de notre système de telle manière que soient mieux distingués ce qu'il est possible d'interpréter étant donné les informations utilisées et ce qui ne l'est pas. L'objectif sera de définir un système qui intègre plus nettement les deux modules suivants [33] :

[33] Le système de Baldwin (voir section 6.5.5) est construit selon le principe défendu ici, mais il nous semble que la précision qu'il atteint (92 %) n'est pas suffisante.

A. un ensemble de règles qui produit pour chaque expression pronominale visée une ou plusieurs réponses, mais la réponse correcte est toujours dans cet ensemble ;
B. un ensemble de préférences qui produit pour chaque expression pronominale une réponse unique mais celle-ci n'est pas garantie correcte.

Cette modularité était plus ou moins visée par nous au départ de l'implantation, mais nos règles se sont révélées insuffisamment fiables, en particulier les règles sur les zones d'antécédence. L'évolution du système devrait donc passer par les opérations suivantes :

- l'ensemble des règles, en particulier les règles sur les zones d'antécédence, devra être corrigé et complété pour permettre de trouver au moins un antécédent correct pour chaque expression pronominale visée (c'est-à-dire améliorer le rappel) ;
- certaines préférences (les préférences 1 à 5 et la préférence 7) semblent avoir une validité forte. Cette validité devra être testée à plus grande échelle, et, si elle se confirme, les préférences en question pourront être adjointes aux règles sans entraîner une baisse significative des réponses correctes.

Ces modifications conduiront à la définition du module A. On notera qu'elles pourront être l'occasion d'examiner la possibilité de formulation plus générale des règles sur les zones d'antécédence et des contraintes.

En ce qui concerne les préférences (module B), un travail de recherche sera à effectuer pour trouver de nouvelles préférences qui expriment des conditions suffisantes pour éliminer correctement des antécédents possibles. Dans cette perspective, le point qui nous semble le plus important sera de prendre en compte de nouvelles sources d'information, question que nous discutons dans la section suivante.

Ajout de nouvelles sources d'information

Nous avons vu au chapitre 6 que l'interprétation des expressions pronominales met en jeu certaines sources d'informations dont nous n'avons pu disposer pour l'implantation décrite dans la seconde partie de la thèse. Parmi ces sources d'informations, les restrictions de sélection (voir section 6.2) constituent la principale source d'information qu'il sera intéressant pour nous d'explorer dans l'avenir.

Les restrictions de sélection ont rarement été utilisées de manière effective dans un système de résolution automatique des expressions pronominales. Les expériences qui ont été menées par Dagan et Itai [26] et Nasukawa [65] (voir section 6.2.2) mettent en jeu des patrons de cooccurrence de termes et non à proprement parler des restrictions de sélection, entendues comme des restrictions formulées pour des classes de termes et non des termes singuliers. Il serait donc

intéressant de tester un système d'hypothèses qui prenne en compte les restrictions de sélection.

Parmi les sources d'information non utilisées dans notre système, on notera que l'information de nature pragmatique décrite dans la section 6.3 est beaucoup plus difficile à modéliser et semble hors de portée des techniques actuelles pour une application sur des textes effectifs.

Élargissement de l'ensemble des expressions pronominales visées

Les reprises visées par notre système ne sont qu'un échantillon des expressions pronominales qu'on rencontre dans les textes. Nous avons dit (p. 171) que notre objectif était de définir dans un premier temps un système qui rende compte de l'interprétation d'un ensemble restreint d'expressions pronominales, puis, dans un second temps, de tester dans quelle mesure le système pourrait rendre compte de l'interprétation d'autres expressions pronominales, telles que les pronoms numéraux (p. ex. *trois*) ou les pronoms démonstratifs (p. ex. *celui-ci*).

Abstraction faite :

- du changement nécessaire dans la représentation de la relation entre le pronom et son antécédent pour les pronoms numéraux et démonstratifs simples (p. ex. *celles* dans *celles de l'Atlantique*) qui ne sont pas interprétés avec une relation de coréférence,
- et des pronoms pouvant renvoyer à une phrase ou proposition (p. ex. *ceci*), dont il est évident que notre système ne peut rendre compte sans ajout de nouvelles règles,

l'idée serait de voir lesquelles de nos règles et préférences sont également valides pour ces nouveaux pronoms. Par exemple, on pourrait faire l'hypothèse que les règles sur les expressions dénotantes ou sur les zones d'antécédence ont la même validité pour les expressions visées à l'heure actuelle et les autres expressions pronominales. Inversement, certaines contraintes ou préférences risquent de ne pas être pertinentes. C'est le cas des contraintes d'accord pour les pronoms démonstratifs simples [34], comme on le voit avec les deux pronoms démonstratifs de la phrase suivante :

(47) Il n'existe pas sur l'euromarché de calendrier comparable à ceux autour desquels s'organise l'activité de nombreux marchés nationaux, comme celui du franc français par exemple.

L'exemple (47) illustre également le fait qu'il faudra sans doute prendre en compte d'autres structures, en l'occurrence les structures exprimant une comparaison, pour rendre compte de l'interprétation des pronoms démonstratifs simples.

[34] On distingue les pronoms démonstratifs simples et composés comme le fait Grevisse [37, §667].

Autre exemple, avec un type d'expression pronominale — les pronoms démonstratifs composés — *a priori* plus proche des expressions visées par notre système (ces formes sont plus proches parce qu'interprétées avec une relation de coréférence). Dans la phrase suivante, le pronom *celui-ci* doit être interprété comme coréférent avec *le gouvernement* :

(48) François Mitterrand a renoué avec l'habitude qu'il avait prise, au cours de la première cohabitation, de « marquer » le gouvernement en faisant connaître, à mesure que celui-ci met en oeuvre sa politique, son désaccord avec les mesures annoncées.

Si on appliquait pour ce pronom les mêmes règles et préférences que pour un pronom *il* dans notre système, on obtiendrait en sortie de la préférence 12, les trois relations suivantes :

```
coref(celui-ci,François Mitterrand)
coref(celui-ci,il)
coref(celui-ci,gouvernement)
```

L'application de la préférence 13 (préférence pour la cohésion du discours) conduirait à retenir la seconde de ces trois relations, et au final, par transitivité étant donné une relation `coref(il,François Mitterrand)`, une interprétation du pronom *celui-ci* comme dénotant François Mitterrand. Nous avons vu que la préférence 13 n'était pas absolument valide pour les expressions visées par le système [35], mais elle semble tout à fait injustifiée pour les démonstratifs composés.

En revanche, les pronoms démonstratifs composés se distinguent des pronoms personnels par le fait qu'ils fonctionnent souvent dans les textes comme des « mentionnels » (suivant le terme de Corblin [24]), c'est-à-dire comme « renvoyant à l'espace des entités discursives [c'est-à-dire les expressions] du texte et à leur ordre relatif [24, p. 35] » pour leur interprétation. Dans (48), l'ambiguïté d'interprétation pour le pronom *celui-ci* en sortie de la préférence 12 pourrait être levée sur la base du fait que le sens de *celui-ci* est de faire référence à l'être mentionné par l'expression qui est la plus proche à gauche de *celui-ci*, c'est-à-dire le gouvernement.

Corblin remarque qu'« une forme comme *celui-ci*, interprétée par reprise immédiate, est en fait beaucoup plus « sûre » qu'un vrai pronom pour ce qui relève de l'identification des référents visés par le locuteur. Il s'agit d'ailleurs d'une propriété distinctive et pour ainsi dire essentielle des mentionnels : contrairement aux pronoms, ils ne sont (quasiment) jamais ambigus. [24, p. 40] » L'intégration

[35] À cet égard, on remarquera que si on remplaçait le pronom *celui-ci* par *il* dans (48),

(48) François Mitterrand a renoué avec l'habitude qu'il_1 avait prise, au cours de la première cohabitation, de « marquer » le gouvernement en faisant connaître, à mesure qu'il_2 met en oeuvre sa politique, son désaccord avec les mesures annoncées.

un observateur quelconque interpréterait toujours le nouveau pronom il_2 comme dénotant le gouvernement, alors que notre système donnerait une réponse incorrecte.

de formules rendant compte de l'interprétation de nouvelles formes pronominales dans notre système pourra permettre de montrer ce que différentes formes ont en commun (par exemple on peut faire l'hypothèse que les règles et les premières préférences que nous avons définies sont valides aussi bien pour les pronoms démonstratifs composés que pour les pronoms personnels) et ce qui les distingue (par exemple, une moindre ambiguïté pour les démonstratifs composés).

Parmi les expressions pronominales qui ne sont pas traitées par notre système, rappelons que figurent également les expressions à sources multiples. Pour traiter ces expressions, un mécanisme de représentation des ensembles d'êtres dénotés chacun par une expression différente devra être intégré dans le système XIP. Nous reviendrons sur ce point dans la section « Problèmes d'implantation » ci-dessous.

Élargissement à l'ensemble des phénomènes de reprise

Au-delà des expressions pronominales évoquées dans la section précédente, nous aimerions développer un système qui rend compte de l'ensemble des phénomènes de reprise, tels que décrits au chapitre 2.

À supposer que la notion de reprise soit opérationnelle (ce nous aurons à démontrer), cette notion est susceptible de conduire à la définition d'une tâche générale pour un système d'interprétation des textes. Cette tâche consisterait, étant donné un texte, :

(i) à identifier les différentes chaînes de référence du texte (une chaîne de référence étant l'ensemble, éventuellement singleton, des expressions qui dans le texte désignent un même être $\mathtt{o}_i$),

(ii) à regrouper ces chaînes de référence en divers ensembles E, chaque ensemble E_i de chaînes de référence étant caractérisé par une description d s'appliquant à chacun des référents associés aux chaînes de référence de E_i,

(iii) à rendre compte des différentes relations de type `membre-de` qui peuvent être observées entre les différents êtres associés aux chaînes de référence.

Le point (i) est une tâche qui est classique depuis le *Coreference Task Definition* de MUC-6 [38]. Le point (iii) ne nécessite pas plus ample discussion ; signalons simplement que c'est une tâche qui à notre connaissance n'a pas été abordée dans un système effectif de traitement automatique des langues.

Le point (ii) viserait à rendre compte des reprises avec identité de descriptions mettant en jeu des expressions dénotantes et ne mettant pas en jeu une relation `membre-de` (cette dernière relevant du point (iii)). On notera que la réalisation de cette tâche impliquera de résoudre certains phénomènes d'anaphore non traités par notre système (par exemple, les reprises de type *les plages de l'Atlantique... celles de la Méditerranée*), mais le point (ii) ne se limite pas cela. Il s'agit de se donner un prédicat d'observation des textes qui exprime une identité dans le *type* des êtres dénotés par les expressions (identité de description), de manière

analogue au prédicat d'observation qui exprime une identité de dénotation entre expressions (coréférence).

Le *Named Entity Task* des conférences MUC [38] peut être vu comme un cas particulier de la tâche exposée dans notre point (ii) : il s'agit d'identifier les noms propres apparaissant dans les textes et de les classer selon qu'ils font référence à une personne, une société ou un lieu, ce qui en un sens revient à classer les référents de ces expressions en fonction de leur type. La tâche est-elle généralisable de manière opérationnelle à un classement de tous les référents évoqués dans un texte sans définition *a priori* des classes d'êtres qui doivent être considérées ? Nous avons pour projet de répondre à cette question.

Problèmes d'implantation

Nous avons évoqués dans le cours de la présentation de notre système certaines limites du système XIP dans lequel il a été implanté. Nous revenons ici sur ces limites en envisageant les développements qui seraient nécessaires.

De manière générale, il serait utile de doter le système XIP d'un mécanisme de représentation des référents, dans le style, par exemple, des « file cards » de Baldwin [6, 7]. Ce mécanisme présenterait deux intérêts principaux. D'une part, il devrait permettre de représenter les référents qui sont des ensembles dont les éléments sont dénotés par des expressions distinctes, à partir de quoi on pourra traiter les reprises avec source multiples. D'autre part, en permettant une représentation des classes d'équivalence que sont les chaînes de coréférence (ou les sous-chaînes de coréférence), il devrait permettre :

- d'éviter d'avoir à gérer de fausses ambiguïtés,
- d'appliquer les certaines contraintes de non-coréférence de manière globale,
- de factoriser au niveau de l'objet informatique représentant un référent certains des traits pertinents pour la résolution des expressions pronominales.

Nous explicitons ici ces trois points. L'exemple (49) suivant illustrera les fausses ambiguïtés et l'application globale des contraintes.

(49) Les principales banques sud-coréennes, emmenées par la Commercial Bank of Korea, ont indiqué hier qu'elles$_1$ décideraient d'ici la fin du mois de juin du sort qu'elles$_2$ réserveront à des dizaines d'entreprises du pays, en les classant selon trois types de catégories : normal, sauvable et non viable.

En sortie des règles, on a pour cette phrase les six relations `coref` suivantes :

(i) `coref(elles`$_1$`,banques)`
(ii) `coref(elles`$_2$`,banques)`
(iii) `coref(elles`$_2$`,elles`$_1$`)`
(iv) `coref(les,banques)`

(v) `coref(les,elles`$_1$`)`
(vi) `coref(les,dizaines)`

et la relation `non-coref` suivante [36] :

`non-coref(les,elles`$_2$`)`

L'existence des deux relations `coref` (ii) et (iii) est ce qu'on appelle une fausse ambiguïté. Dans la mesure où le pronom *elles*$_1$ est non ambigu pour le système, que *elles*$_2$ renvoie à *elles*$_1$ ou à l'antécédent de *elles*$_1$ n'exprime pas une réelle ambiguïté sur l'antécédent de *elles*$_2$.

Les trois premières relations `coref` définissent donc la classe d'équivalence suivante :

`{banques, elles`$_1$`, elles`$_2$`}`

Étant donné cette classe d'équivalence, la contrainte de non-coréférence exprimée entre les deux expressions *les* et *elles*$_2$ devrait automatiquement s'appliquer entre *les* et *elles*$_1$ et *les* et *banques*. En d'autres termes, l'ensemble des relations qu'on a pour l'exemple (49) en sortie des règles exprime en fait un résultat non ambigu. Il serait bon que le système dispose de mécanismes qui fassent apparaître plus clairement cette non-ambiguïté.

Notre troisième idée de développement est de factoriser au niveau de l'objet informatique représentant un référent un certain nombre d'informations sémantiques sur ce référent.

Supposons que pour interpréter le pronom *les* dans le texte suivant, on souhaite faire usage de restrictions de sélection, qui disent, par exemple, que l'objet du verbe *amener* suivi d'une infinitive dénote de préférence une personne.

(50) Les équipes du ministère des Finances en charge des dossiers financiers ne chôment pas en ce printemps. Tout en gérant de front les négociations épiques avec Bruxelles sur le Crédit Lyonnais et les privatisations du GAN, du Crédit Foncier et de la Société Marseillaise de Crédit, elles planchent également sur le dossier sensible et lourd de la réforme des Caisses d'Epargne, avec pour objectif de présenter un projet de loi à la fin du mois de juin. Ce calendrier serré les a amenées à se pencher sur le problème des structures de direction de l'Ecureuil, dont les mandats arrivent à échéance également fin juin.

Parmi les antécédents possibles pour *les* en sortie des règles figure l'antécédent correct *elles*. En pratique, dans notre système, on n'a pas d'information sur le type de l'être dénoté par *elles* et il serait donc impossible de faire usage de

[36] Cette relation est identifiée par la règle C-R.1 parce que l'expression *elles*$_2$ est identifiée par l'analyseur syntaxique comme celle qui contrôle le sujet du participe présent *classant*. Une analyse peut-être plus juste aurait été de donner ce rôle à l'expression *elles*$_1$, mais cela n'a pas d'incidence sur notre propos présent.

la restriction de sélection que nous avons évoquée. Par contre, si on avait une représentation de la classe d'équivalence

```
{équipes, elles}
```

équipes étant l'antécédent préalablement identifié pour le pronom *elles*, alors on pourrait associer à cette classe d'équivalence le trait `person:+`, à partir de l'information disponible au niveau de l'unité lexicale *équipes* et la restriction de sélection pourrait alors être utilisée.

Notons que le processus décrit ici implique idéalement une interprétation du texte phrase à phrase, plutôt que le mécanisme actuel, où chaque règle est appliquée tour à tour sur l'ensemble du texte. L'implantation de ce mécanisme d'analyse phrase à phrase, corollaire de l'implantation d'un mécanisme de représentation des référents, devra être étudiée.

Bibliographie

[1] Salah Aït-Mokhtar et Jean-Pierre Chanod. Incremental finite-state parsing. In *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP-97)*, Washington, DC, USA, 1997.

[2] Salah Aït-Mokhtar et Jean-Pierre Chanod. Subject and object dependency extraction using finite-state transducers. In *ACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, Madrid, 1997.

[3] Salah Aït-Mokhtar, Jean-Pierre Chanod et Claude Roux. A multi-input dependency parser. In *Proceedings of the 7th International Workshop on Parsing Technologies (IWPT-2001)*, Beijing, 17-19 octobre 2001.

[4] Chinatsu Aone et Douglas McKee. A language-independent anaphora resolution system for understanding multilingual texts. In *Proceedings of ACL'93*, p. 156–163, Ohio State University, États-Unis, 1993.

[5] Amit Bagga et Breck Baldwin. Algorithms for scoring coreference chains. In *Proceedings of the LREC'98 Workshop on Linguistic Coreference*, Grenade, Espagne, 1998.

[6] Breck Baldwin. *CogNIAC : A Discourse Processing Engine*. Thèse de doctorat, Faculty of Engineering and Applied Science, University of Pennsylvania, 1995.

[7] Breck Baldwin. CogNIAC : High precision coreference with limited knowledge and linguistic resources. In *Proceedings of the ACL'97/EACL'97 Workshop on Operational factors in practical, robust anaphora resolution*, p. 38–45, Madrid, Espagne, 1997.

[8] Cătălina Barbu et Ruslan Mitkov. Evaluation tool for rule-based anaphora resolution methods. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL'2001)*, p. 34–41, Toulouse, France, 2001.

[9] Michael Barlow. Feature mismatches and anaphora resolution. In *Proceedings of the Second Colloquium on Discourse Anaphora and Anaphor Resolution (DAARC2)*, p. 34–42, Lancaster University, Royaume-Uni, 1998.

[10] Gabriel G. Bès. La phrase verbale noyau en français. In *Recherches sur le français parlé*, p. 237–358. Publications de l'Université de Provence, 1999.

[11] Gabriel G. Bès. Empiricité en linguistique et grammaire de Montague : la sémantique en 5P et la compositionalité. GRIL, Université Blaise-Pascal, avril 2001.

[12] Simon Botley et Anthony M. McEnery, directeurs de publication. *Corpus-based and Computational Approaches to Discourse Anaphora*, chapitre Discourse Anaphora : The need for synthesis. John Benjamins, Amsterdam, 2000.

[13] Susan E. Brennan, Marilyn W. Friedman et Carl J. Pollard. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics (ACL'87)*, p. 155–162, Stanford, Californie, États-Unis, 1987.

[14] Donna K. Byron. The Uncommon Denominator. *Computational Linguistics - Special Issue on Anaphora Resolution*, Á paraître. Disponible à l'adresse : `http://www.cs.rochester.edu/u/dbyron/papers.html`.

[15] Donna K. Byron et Joel R. Tetreault. A flexible architecture for reference resolution. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL'99)*, p. 229–232, Bergen, Norvège, 1999.

[16] Jaime G. Carbonnel et Ralf G. Brown. Anaphora resolution : A multistrategy approach. In *Proceedings of the 12th International Conference on Computational Linguistics (COLING'88)*, p. 96–101, Budapest, 1998.

[17] David M. Carter. *Interpreting anaphors in natural language texts.* Ellis Horwood, Royaume-Uni, 1987.

[18] Michel Chambreuil et Jean-Claude Pariente. *Langue naturelle et logique : La sémantique intensionnelle de Richard Montague.* Peter Lang, Berne, 1990.

[19] Michel Charolles et Catherine Schnedeker. Coréférence et identité. Le problème des référents évolutifs. *Langages*, 112, p. 106–126, 1993.

[20] Nancy A. Chinchor et Beth Sundheim. Message Understanding Conference (MUC) Tests of Discourse Processing. In *Proceedings of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, Stanford, 1995.

[21] Noam Chomsky. *Lectures on Government and Binding.* Foris Publications, Dordrecht, 1981.

[22] Dennis Connolly, John D. Burger et David S. Day. A machine learning approach to anaphoric reference. In Daniel B. Jones et Harold L. Somers, directeurs de publication, *New Methods in Language Processing*, p. 133–144. UCL Press, Londres, 1997.

[23] Francis Corblin. *Les formes de reprise dans le discours. Anaphore et chaînes de référence.* Presses Universitaires de Rennes, 1995.

[24] Francis Corblin. *Celui-ci* anaphorique : un mentionnel. *Langue française*, 120, p. 33–43, 1998.

[25] Dan Cristea, Nancy Ide et Laurent Romary. Veins theory : A model of global discourse cohesion and coherence. In *Proceedings of COLING-ACL'98*, p. 281–285, Montréal, Canada, 1998.

[26] Ido Dagan et Alon Itai. Automatic processing of large corpora for the resolution of anaphora references. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING'90)*, p. 330–332, Helsinki, 1990.

[27] Laurence Danlos. Event coreference in causal discourses. In Pierrette Bouillon et Federica Busa, directeurs de publication, *The Language of Word Meaning*, p. 216–241. Cambridge University Press, 2001.

[28] Sarah Davies et Massimo Poesio. Coding Schemes for Co-reference. `http://www.cogsci.ed.ac.uk/~poesio/MATE/coreference.html`, 1998.

[29] Barbara Di Eugenio. On the usage of kappa to evaluate agreement on coding tasks. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athènes, Grèce, 2000.

[30] David R. Dowty, Robert E. Wall et Stanley Peters. *Introduction to Montague Semantics.* D. Reidel Publishing Company, Dordrecht, Holland, 1981.

[31] Oswald Ducrot et Tzvetan Todorov. *Dictionnaire encyclopédique des sciences du langage.* Éditions du Seuil, Paris, 1972.

[32] Antonio Ferrández, Manuel Palomar et Lidia Moreno. Anaphora resolution in unrestricted texts with partial parsing. In *Proceedings of COLING-ACL'98*, p. 385–391, Montréal, Canada, 1998.

[33] Steve Fligelstone. Developing a scheme for annotating text to show anaphoric relations. In G. Leitner, directeur de publication, *New Directions in Corpus Linguistics*, p. 153–170. Mouton de Gruyter, Berlin, 1992.

[34] Gottlob Frege. *Écrits logiques et philosophiques.* Éditions du Seuil, Paris, 1971.

[35] Niyu Ge, John Hale et Eugene Charniak. A statistical approach to anaphora resolution. In *Proceedings of the 6th Workshop on Very Large Corpora*, p. 161–170, Montréal, Canada, 1998.

[36] Gilles-Gaston Granger. *La vérification.* Éditions Odile Jacob, Paris, 1992.

[37] Maurice Grevisse. *Le bon usage. Grammaire française. Refondue par André Goosse.* Duculot, Paris - Louvain-la-Neuve, treizième édition, 1993.

[38] Ralph Grishman et Beth Sundheim, directeurs de publication. *Proceedings of the Sixth Message Understanding Conference (MUC-6).* Morgan Kaufmann, San Francisco, 1995.

[39] Barbara J. Grosz, Aravind K. Joshi et Scott Weinstein. Centering, a framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2), p. 175–204, 1995.

[40] Barbara J. Grosz et Candace L. Sidner. Lost intuitions and forgotten intentions. In Marilyn A. Walker, Aravind K. Joshi et Ellen F. Prince, directeurs de publication, *Centering Theory in Discourse*, p. 39–51. Oxford University Press, New York, 1998.

[41] Gsi-Erli. Le dictionnaire AlethDic. Version 1.5.4, 15 décembre 1994.

[42] Franz Günthner et Hubert Lehmann. Rules for pronominalisation. In *Proceedings of the First Conference of the European Chapter of the Association for Computational Linguistics*, p. 144–151, Pise, Italie, 1983.

[43] M. A. K. Halliday et Ruqaiya Hasan. *Cohesion in English.* Longman, 1976.

[44] Lynette Hirshman et Nancy Chinchor. MUC-7 Coreference Task Definition. Version 3.0. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, `http://www.muc.saic.com`, 1998. Science Applications International Corporation.

[45] Jerry Hobbs. Resolving pronoun references. *Lingua*, 44, p. 311–338, 1976.

[46] Nancy Ide et Dan Cristea. A hierarchical account of referential accessibility. In *Proceedings of ACL-2000*, p. 416–424, Hong Kong, 2000.

[47] Kerstin Jonasson. *Le nom propre. Constructions et interprétations.* Duculot, Louvain-la-Neuve, 1994.

[48] Megumi Kameyama. Intrasential centering : A case study. In Marilyn A. Walker, Aravind K. Joshi et Ellen F. Prince, directeurs de publication, *Centering Theory in Discourse*, p. 89–112. Oxford University Press, New York, 1998.

[49] Hans Kamp, Dick Crouch et Josef van Genabith. A Framework for Computational Semantics (FRACAS). Deliverable D2. Specification of Linguistic Coverage. `http://www.cogsci.ed.ac.uk/~fracas/`, 1994.

[50] Hans Kamp et Uwe Reyle. *From Discourse to Logic.* Kluwer Academic Publishers, 1993.

[51] Lauri Karttunen. Discourse referents. In James D. McCawley, directeur de publication, *Syntax and Semantics. Notes from the Linguistic Underground*, p. 363–385. Academic Press, New York, 1976.

[52] Andrew Kehler, John Bear et Douglas Appelt. The Need for Accurate Alignment in Natural Language System Evaluation. *Computational Linguistics*, 27(2), 2001.

[53] Christopher Kennedy et Branimir Boguraev. Anaphora for everyone : Pronominal anaphora resolution without a parser. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, Copenhague, Danemark, 1996.

[54] Georges Kleiber, Catherine Schnedeker et Laurence Ujma. L'anaphore associative, d'une conception l'autre. In *L'anaphore associative. Recherches linguistiques XIX*, p. 5–64. Université de Metz, 1994.

[55] Shalom Lappin et Herbert J. Leass. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4), p. 535–561, 1994.

[56] William C. Mann et Sandra A. Thompson. Rhethorical structure theory : Toward a functional theory of text organization. *Text*, 8(3), p. 243–281, 1988.

[57] Daniel Marcu, Estibaliz Amorrortu et Magdalena Romera. Experiments in constructing a corpus of discourse trees. In *Proceedings of the ACL'99 Workshop on Standards and Tools for Discourse Tagging*, p. 48–57, Maryland, États-Unis, 1999.

[58] Patricio Martínez-Barco, Rafael Muñoz, Saliha Azzam, Manuel Palomar et Antonio Ferrández. Evaluation of pronoun resolution algorithm for spanish dialogues. In *Proceedings of VEXTAL'99*, p. 325–332, Università Ca' Foscari, Venise, Italie, 1999.

[59] MATE – Multilevel Annotation, Tools Engineering. Telematics Project LE4-8370. `http://mate.nis.sdu.dk`, 2000.

[60] Ruslan Mitkov. Robust pronoun resolution with limited knowledge. In *Proceedings of COLING-ACL'98*, p. 869–875, Montréal, Canada, 1998.

[61] Ruslan Mitkov. Anaphora resolution : The state of the art. Working paper (Based on the COLING/ACL'98 tutorial on anaphora resolution). `http://www.wlv.ac.uk/sles/compling/papers/mitkov-99a.pdf`, University of Wolverhampton, Royaume-Uni, 1999.

[62] Ruslan Mitkov. Towards more comprehensive evaluation in anaphora resolution. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, p. 1309–1314, Athènes, Grèce, 2000.

[63] Ruslan Mitkov, Lamia Belguith et Malgorzata Stys. Multilingual robust anaphora resolution. In *Proceedings of the Third International Conference on Empirical Methods in Natural Language Processing (EMNLP-3)*, p. 7–16, Grenade, Espagne, 1998.

[64] Fiammetta Namer. *Pronominalisation et effacement du sujet en génération automatique de textes en langues romanes.* Thèse de doctorat, Université Paris VII, 1990.

[65] Tetsuya Nasukama. Robust method of pronoun resolution using full-text information. In *Proceedings of the 15th Conference on Computational Linguistics (COLING'94)*, p. 1157–1163, Kyoto, Japon, 1994.

[66] Constantin Orăsan, Richard Evans et Ruslan Mitkov. Enhancing preference-based anaphora resolution with genetic algorithms. In *Proceedings of the Second Natural Language Processing International Conference (NLP-2000)*, p. 185–195, Patras, Grèce, 2000.

[67] Rebecca Passonneau. Applying reliability metrics to co-reference annotation. Rapport Technique CUCS-017-97, Columbia University, Department of Computer Science, 1997.

[68] Jesús Peral, Manuel Palomar et Antonio Ferrández. Coreference-oriented interlingual slot structure and machine translation. In *Proceedings of the ACL Workshop on Coreference and Its Applications*, p. 69–76, University of Maryland, États-Unis, 1999.

[69] Massimo Poesio. MATE Annotation Guidelines – Coreference. 2000. `http://www.ims.uni-stuttgart.de/projekte/mate/mdag/cr/cr_1.html`.

[70] Massimo Poesio et Renata Vieira. A corpus-based investigation of definite descritpion use. *Computational Linguistics*, 24(2), p. 183–216, 1998.

[71] Andrei Popescu-Belis. Évaluation numérique de la résolution de la référence : critiques et propositions. *Traitement Automatique des Langues*, 40(2), p. 117–142, 1999.

[72] Andrei Popescu-Belis. *Modélisation multi-agent des échanges langagiers : application au problème de la référence et à son évaluation.* Thèse de doctorat, Université Paris XI, Orsay, 1999.

[73] Andrei Popescu-Belis et Isabelle Robba. Cooperation between pronoun and reference resolution for unrestricted texts. In *Proceedings of the ACL'97/EACL'97 Workshop on Operational factors in practical, robust anaphora resolution*, p. 94–99, Madrid, Espagne, 1997.

[74] Andrei Popescu-Belis et Isabelle Robba. Three new methods for evaluating reference resolution. In *Proceedings of the LREC'98 Workshop on Linguistic Coreference*, Granada, Spain, 1998.

[75] Aarne Ranta. *Type-Theoretical Grammar.* Oxford University Press, Oxford, 1994.

[76] Anne Reboul, Cécile Balkanski, Xavier Briffault, Bertrand Gaiffe, Andrei Popescu-Belis, Isabelle Robba, Laurent Romary et Gérard Sabah. Le projet CERVICAL : Représentations mentales, référence aux objets et aux événements. Rapport technique, CRIN et LIMSI, 1997.

[77] Tanya Reinhart. *Anaphora and Semantic Interpretation.* Croom Helm, Londres, 1983.

[78] Tanya Reinhart. Coreference and bound anaphora : A restatement of the anaphora questions. *Linguistics and Philosophy*, 6(1), p. 47–88, 1983.

[79] Monique Rolbert. Résolution de formes pronominales dans l'interface d'interrogation d'une base de données. Thèse de doctorat, Faculté des sciences de Luminy., 1989.

[80] Maximiliano Saiz-Noeda et Manuel Palomar. Semantic knowledge-driven method to solve pronominal anaphora in spanish texts. In *Natural Language Processing — NLP 2000, Second International Conference*, p. 204–221, Patras, Grèce, 2000. Springer.

[81] Susanne Salmon-Alt. Du corpus à la théorie : l'annotation (co-)référentielle. *Traitement Automatique des Langues*, 42(2), 2001.

[82] Susanne Salmon-Alt. *Référence et dialogue finalisé : de la linguistique à un modèle opérationnel.* Thèse de doctorat, Université Henri-Poincaré, Nancy 1, 2001.

[83] Candace L. Sidner. Focusing for interpretation of pronouns. *Computational Linguistics*, 7(4), p. 217–231, 1981.

[84] Michael Strube et Udo Hahn. Functional centering — Grounding referential coherence in information structure. *Computational Linguistics*, 25(3), p. 309–344, 1999.

[85] Roland Stuckardt. Robust anaphor resolution : Design and evaluation of the ROSANA system. In *Proceedings of ROMAND 2000 : Workshop on Robust Methods in Analysis of Natural Language Data*, p. 43–57, Lausanne, 2000.

[86] Richmond H. Thomason, directeur de publication. *Selected papers of Richard Montague.* Yale University Press, New Haven, 1974.

[87] François Trouilleux. Identification et classement automatique des noms propres dans des textes en français. Mémoire de DEA, GRIL, Université Blaise-Pascal, Clermont-Ferrand, 1997.

[88] François Trouilleux, Gabriel G. Bès et Éric Gaussier. An evaluation of inter-annotator agreement in the observation of anaphoric and referential relations. In *Proceedings of the Third International Conference on Discourse Anaphora and Anaphor Resolution (DAARC2000)*, Lancaster University, Royaume-Uni, 2000.

[89] François Trouilleux, Éric Gaussier, Gabriel Bès et Annie Zaenen. Coreference resolution evaluation based on descriptive specificity. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athènes, Grèce, 2000.

[90] Agnès Tutin, François Trouilleux, Catherine Clouzot, Éric Gaussier, Annie Zaenen, Stéphanie Rayot et Georges Antoniadis.

Annotating a large corpus with anaphoric links. In *Proceedings of the Third International Conference on Discourse Anaphora and Anaphor Resolution (DAARC2000)*, Lancaster University, Royaume-Uni, 2000.

[91] Marc Vilain, John Burger, John Aberdeen, Dennis Connolly et Lynette Hirshman. A model-theoretic coreference scoring scheme. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, p. 45–52, San Francisco, 1995. Morgan Kaufmann.

[92] Marilyn A. Walker. Evaluating discourse processing algorithms. In *Proceedings of ACL'89*, p. 251–261, University of British Columbia, Canada, 1989.

[93] Marilyn A. Walker, Aravind K. Joshi et Ellen F. Prince, directeurs de publication. *Centering Theory in Discourse.* Oxford University Press, New York, 1998.

[94] Yorick Wilks. *Preference Semantics.* Stanford AI Laboratory memo AIM-206, Stanford University, 1973.

[95] Terry Winograd. *Understanding Natural Language.* Academic Press, New York, 1972.

[96] Anne Zribi-Hertz. *L'anaphore et les pronoms. Une introduction à la syntaxe générative.* Presses Universitaires du Septentrion, Villeneuve-d'Ascq, 1996.

Annexes

Annexe A

Données du test d'opérationalité

Cette annexe présente les trois textes utilisés pour le test d'opérationalité décrit au chapitre 4, avec l'annotation de référence, ou annotation clé, réalisée par l'expert (section A.1), puis, sous forme de tableaux, l'ensemble des liens observés par les six participants à l'expérience, avec les prédicats d'évaluation pour chacune des observations faites par les cinq annotateurs, ainsi que l'opinion majoritaire qui se dégage de ces cinq annotations, cette dernière constituant l'annotation de l'« observateur idéal »

Les textes présentés ici sont annotés avec le schéma d'annotation présenté au chapitre 2. Des indices ont été ajoutés pour permettre un renvoi des textes aux tableaux présentés plus loin (section A.7). Dans l'annotation de référence, un nombre en indice identifie un lien particulier ; ce nombre est repris dans les tableaux de la section A.7. Un X en indice indique que l'expression est prise comme représentative d'un référent. Dans les annotations réponses, sont placés en indice des symboles exprimant les jugements qui sont faits sur les différentes liens :

- $\emptyset$ note que l'annotation n'est pas prise en compte (voir section 4.1.7) ;
- T note que le lien est jugé explicite dans le texte (T pour « trivial » ; voir la section 4.1.2) ;
- C note que l'annotation est jugée correcte ;
- I note que l'annotation est jugée incorrecte ;
- $X - Y$, où X et Y sont des symboles identifiant un type de lien caractérise un lien noté comme étant de type X dans la réponse alors qu'il est de type Y dans la clé ; pour un inventaire des différents symboles possibles voir la section A.7 ;
- enfin la lettre « s » suivie d'un nombre caractérise les liens jugés superflus.

Trois points sont à noter. Les liens sont vus ici comme orientés suivant le critère que nous avons défini pour les liens de reprise (voir p. 29). La notation `{o`$_j$, ..., `o`$_n$`}-id-o`$_i$ n'était pas prévue dans le schéma d'annotation donnée aux

étudiants. Ceux-ci devaient utiliser la notation {o$_j$, ..., o$_n$}-mde-o$_i$, signifiant que l'extension de l'ensemble o$_i$ est entièrement spécifié par l'énumération à gauche de la suite -mde-. Ils n'ont pas toujours respecté cette directive et on utilisé cette notation dans des cas où elle n'était pas prévue. Nous avons donc, dans ces cas, décomposé leur notation en autant de liens que d'être spécifiés à gauche de -mde-.

Enfin, signalons que, mis à part le formatage des fichiers originaux pour l'inclusion dans la thèse et l'ajout des indices, les annotations réponses sont livrées sans aucune modification.

A.1 Annotation de référence

Texte B

<[vendredi 29 mai 1998] >

<La BNP [o1]$_X$> réorganise < <son [o1]$_1$> pôle [o2]$_X$> de <financements [o6]$_X$> spécialisés...

Comme <elle [o1]$_2$> <l' [o5(P)]$_3$>avait laissé entendre en <février [février 1998]$_4$>, < <la BNP [o1]$_5$> a décidé de rapprocher < <ses [o1]$_6$> filiales [o2]$_7$> de <crédits [o6]$_8$> spécialisés, le Crédit Universel (<biens [o7]$_9$> d'équipement pour les entreprises, les professionnels et les particuliers) et BNP Bail (<crédit-bail [o8]$_{10}$> mobilier et immobilier, <location [o9]$_{11}$>) [o5]$_X$>.

<[o7-mde-o6]$_9$>

<[o8-mde-o6]$_{10}$>

<[o9-mde-o6]$_{11}$>

<Le nouvel ensemble [o2]$_{12}$>, baptisé BNP Lease, affiche, sur la base de 1997, un produit net bancaire de <1,7 milliard [franc]$_X$> de francs et un résultat avant impôts de <700 millions [franc]$_{13}$>.

< <Sa [o2]$_{14}$> production [o3]$_X$> annuelle < <s' [o3]$_{15}$>élève [s'élever]$_X$> à <20 milliards [franc]$_X$> de francs et < <ses [o2]$_{16}$> encours [o4]$_X$> <[<s' [o4]$_\emptyset$>élever]$_{17}$> à <54 milliards [franc]$_{18}$>.

<Le nouvel ensemble [o2]$_{19}$>, détenu à 100 % par <la banque [o1]$_{20}$>, sera présidé par Claude Porcherot, tandis que Jean-René Brunon <en [o2]$_{21}$> assurera la direction générale.

Texte G

<[mardi 26 mai 1998, o0] >

<Guigou [o1]$_X$> visite <les locaux [o2]$_X$> du <pôle [o3]$_X$> financier

Le ban et l'arrière-ban de <la magistrature [o25]$_X$> parisienne étaient présents, <hier [25 mai 1998]$_{22}$>, aux côtés d'<Elisabeth Guigou [o1]$_{23}$> pour visiter <les futurs locaux [o2]$_{24}$> du <pôle [o3]$_{25}$> financier parisien.

Après des mois de discussion, <le choix [o6]$_X$> définitif <s' [o6]$_{26}$>est porté, comme prévu, sur <le siège [o2]$_{27}$> historique du journal le Monde, rue des Italiens (La Tribune du <19 mai [19 mai 1998]$_{28}$>), au coeur du quartier parisien de la finance.

<Cet immeuble [o2]$_{29}$> luxueux, complètement réaménagé, accueillera sur <6.400 mètres [o11]$_X$> carrés, <d'<ici [o0]$_{30}$> à la fin de <l'année [1998]$_{31}$> [o40]$_X$>, 274 <magistrats [o19]$_X$> et <fonctionnaires [o8]$_X$>, plus <une trentaine [o9]$_X$> d'assistants spécialisés des Finances et de la Banque de France.

<[o19-mde-o25]$_{32}$>

<Ils [o10]$_X$> auront à <leur [o10]$_{34}$> disposition <quelque 23 mètres [o11]$_{35}$> carrés <par personne [o10(D)]$_{36}$>, contre pratiquement <moitié [o12]$_X$> moins <auparavant [o41]$_X$> au <Palais [o7]$_X$> de justice.

<[{o19,o8,o9}-mde-o10]$_{33}$>

<[o12-rel-o11]$_{37}$>

<[o41-rel-o40]$_{38}$>

<Montant [o14]$_X$> du <loyer [o13]$_X$> : <21,6 millions [franc, o14]$_{40}$> de francs par an <auxquels [o14]$_{41}$> <s' [o15]$_{42}$>ajoutent <15 millions [franc, o15]$_{43}$> de travaux spécifiques pour sécuriser <les lieux [o2]$_{44}$> <que [o15]$_{45}$> prend en charge le propriétaire de <l'immeuble [o2]$_{46}$>, la Bred.

<[o13-rel-o2]$_{39}$>

Outre <des effectifs [o10]$_{47}$> regroupés et supplémentaires, <le pôle [o3]$_{48}$> financier bénéficiera d'<équipements [o16]$_X$> informatiques.

<Les magistrats [o19]$_{49}$> auront notamment à <leur [o19]$_{50}$> disposition <des logiciels [logiciel, o17]$_X$> d'instruction assistée par ordinateur.

<[o17-mde-o16]$_{51}$>

<Actuellement [o0]$_{52}$>, seuls <les juges [o50]$_X$> Eva Joly et Jean-Pierre Za-

notto <en [logiciel]$_{54}$> disposent à la galerie financière de Paris.

<[o50-mde-o25]$_{53}$>

A <ceux [o18]$_X$> <qui [o18]$_{55}$> <se [o18]$_{56}$> montrent réticents sur la mise en place de <ce pôle [o3]$_{57}$> financier, <le ministre [o1]$_{58}$> de la Justice a tenu à répéter que, « qu'<on [o20]$_X$> <le [o24(P)]$_{59}$> veuille ou non, < <on [o20]$_{60}$> n'échappera pas à un besoin de spécialisation croissante des <magistrats [o21(G)]$_X$> en matière d'information économique et financière [o24]$_X$> ».

<[o19-mde-o21]$_{61}$>

<Cette annexe [o3]$_{62}$> parisienne du <Palais [o7]$_{63}$> de justice dédiée aux dossiers financiers devrait rapidement être suivie d'<autres pôles [o4]$_X$> en province.

<[o4-dde-o3]$_{64}$>

<Le premier [o5]$_X$> sur <la liste [o22]$_X$> du <gouvernement [o23]$_X$> est le pôle corse.

<[o5-mde-o4]$_{65}$>
<[o22-rel-o4]$_{66}$>
<[o1-mde-o23]$_{67}$>

Texte A

<[francais, o48]>
<[vendredi 29 mai 1998, o0]>

Fort du <rachat [o8]$_X$> des <AGF [o2]$_X$>, <Allianz [o1]$_X$> présente <son [o1]$_{68}$> nouveau visage

+ <L'assureur [o1]$_{69}$> allemand a engagé un tournant stratégique avec <le rachat [o8]$_{70}$> des <AGF [o2]$_{71}$> au début de <l'année [1998, o4]$_{72}$>.

+ <Le développement [o14]$_X$> de <sa [o1]$_{73}$> présence en <France [o24]$_X$> <l' [o1]$_{74}$> amènera < <se [o1]$_{75}$> faire coter à <Paris [o27]$_X$> [o26]$_X$> <le 12 juin [12 juin 1998, o6]$_{76}$> prochain.

<[o14-rel-o8]$_{77}$>

C'est <un nouveau groupe [o1]$_{78}$> Allianz <qui [o1]$_{79}$> naîtra d'<ici [o0]$_{80}$> à <la fin [o7]$_X$> de <l'année [o4]$_{81}$>.

<L'assureur [o1]$_{82}$> allemand, < <qui [o1]$_{83}$> consolide <ses [o1]$_{84}$> 51 % des <AGF [o2]$_{85}$> depuis <le 1er avril [1 avril 1998]$_{88}$> [o8]$_{86}$>, considère que < <cette prise [o8]$_{87}$> de contrôle <lui [o1]$_{89}$> confère « une très forte

position dans <le secteur [o20]$_X$> de l'assurance mondiale, avec un pied particulièrement solide dans < <notre [o1]$_{90}$> marché [o23]$_X$> domestique <qu' [o23]$_{91}$>est l'Europe », [o15]$_X$> comme <l' [o15(P)]$_{92}$> explique < <son [o1]$_{93}$> président [o3]$_X$>, Hennig Schulte-Noelle.

<[o23-mde-o20]$_{94}$>

Tout en <se [o1]$_{95}$> félicitant de < <son [o1]$_{96}$> acquisition [o8]$_{97}$>, <Allianz [o1]$_{98}$> n'en rappelle pas moins < <ses [o1]$_{99}$> objectifs [o28]$_X$>.

<Contre-pied [o25]$_X$> d'<Axa [o9]$_X$>.

< D'<ici [o0]$_{100}$> à 2000, le résultat net des <AGF [o2]$_{101}$> devra être porté à 5,5 milliards de francs, permettant à < <sa [o2]$_{102}$> maison [o1]$_{103}$> mère de constater un retour sur investissement de 9 %. [o29]$_X$>

<[o29-mde-o28]$_{104}$>

Pour 1998, <l'acquisition [o8]$_{105}$> des <AGF [o2]$_{106}$> doit permettre à <Allianz [o1]$_{107}$> de <porter <son [o1]$_{108}$> chiffre d'affaires à <107 milliards [o16]$_X$> de deutsche marks (<358 milliards [o16]$_{109}$> de francs) [o30]$_X$> et de <constater une croissance à deux chiffres de <son [o1]$_{112}$> bénéfice net. [o31]$_X$>

<[o30-mde-o28]$_{110}$>
<[o31-mde-o28]$_{111}$>

Mais pour <l'heure [o0]$_{113}$>, la priorité est à <la « digestion » [o17]$_X$> de <cette opération [o8]$_{115}$> d'<ici [o0]$_{116}$> à fin 1999 - <ce [o17]$_{117}$> <qui [o17]$_{118}$> passe par la détermination, <cette année [o4]$_{119}$>, des structures , des cadres dirigeants et des plans de développement dans chaque pays.

<[o17-mde-o28]$_{114}$>

Au passage, <l'assureur [o1]$_{120}$> allemand rappelle < <sa [o1]$_{121}$> volonté [o32]$_X$> de décentralisation et, comme pour prendre <le contre-pied [o25]$_{123}$> de < <son [o1]$_{124}$> grand concurrent [o9]$_{125}$> Axa, affirme qu'<il [o1]$_{126}$> ne voit pas l'intérêt de coiffer < <ses [o1]$_{127}$> sociétés [o40]$_X$> nationales d'une marque mondiale.

<[o32-mde-o28]$_{122}$>

A <l'heure [o0]$_{128}$> actuelle, aucune décision n'a été prise quant à </- d'éventuelles -/ cessions d'/-<actifs [o37]$_X$>, et notamment -/ <les 25 % [o38]$_X$> de la Coface [o39]$_X$>, comme <le [o39(P)]$_{130}$> demande Bruxelles.

<[o38-mde-o37]$_{129}$>

Face à cette nouvelle donne, <Allianz [o1]$_{131}$> réfléchit à la façon de <parvenir à une présentation commune et aux économies possibles entre <Euler [o41]$_X$>

et <Hermès [o42]$_X$>, sans pour autant créer une structure faîtière. [o33]$_X$>

<[o33-mde-o28]$_{132}$>
<[o41-mde-o40]$_{133}$>
<[o42-mde-o40]$_{134}$>

Enfin, toujours en <France [o24]$_{135}$>, <Hennig Schulte-Noelle [o3]$_{136}$> a réaffirmé <son [o3]$_{137}$> intention d'<entrer jusqu'à 10 % dans le capital du Crédit Lyonnais. [o34]$_X$>

<[o34-mde-o28]$_{138}$>

D'un point de vue stratégique, <cette année [o4]$_{139}$> devrait marquer, selon <Allianz [o1]$_{140}$>, <le démarrage [o35]$_X$> de <sa [o1]$_{142}$> coopération dans la gestion d'actifs avec <la Dresdner Bank [o10]$_X$>, pour <laquelle [o10]$_\emptyset$> la répartition future des compétences reste en négociation.

<[o35-mde-o28]$_{141}$>

En <assurance [o54]$_X$>, <Allianz [o1]$_{143}$> discute avec la nouvelle Bayerische und Vereinsbank <une répartition [o18]$_X$> des partenariats entre <lui [o1]$_{145}$> et <Ergo [o11]$_X$>.

<[o18-mde-o28]$_{144}$>

A <ce sujet [o18]$_{146}$>, et pour tenter de briser le soupçon de <cartel [o43]$_X$> porté sur <le marché [o53]$_X$> allemand, <Allianz [o1]$_{148}$> rappelle que <ses [o1]$_{149}$> 10 % d'<Ergo [o11]$_{150}$> ne sont pas stratégiques et que <les deux compagnies [o12]$_{151}$> restent « des concurrentes acharnées ».

<[o1,o11-mde-o12]$_{151}$>
<[o43-rel-o12]$_{147}$>
<[o53-mde-o23]$_{152}$>

Du reste, <Allianz [o1]$_{153}$> (<qui [o1]$_{154}$> n'a connu en 1997 qu'une hausse de <ses [o1]$_{155}$> primes de 0,4 % sur <le marché [o19]$_X$> automobile allemand en croissance moyenne de 1,3 %), affirme qu'< <il [o1]$_{157}$> « continuera à prendre <des mesures [o44]$_X$> pour maintenir, sinon augmenter, <sa [o1]$_{158}$> part de marché sur <ce segment [o19]$_{159}$> très concurrentiel ». [o36]$_X$>

<[o36-mde-o28]$_{160}$>
<[o19-mde-o53]$_{156}$>

Autant dire que <la guerre [o45]$_X$> des tarifs en Allemagne n'est pas terminée.

<[o44-rel-o45]$_{161}$>

En 1997, <l'assureur [o1]$_{162}$>, <qui [o1]$_{163}$> a publié un bénéfice définitif de <2,7 milliards [mark]> de marks (+ 20,5 %), a, pour la première fois, affiché <l'an [o5]$_{164}$> dernier <un bénéfice [o56]$_X$> technique positif de <182 millions

[mark]$_{165}$>.

Mais en <assurance-dommages [o55]$_X$>, <le chiffre [o57]$_X$> reste négatif de <1,3 milliard [mark]$_{168}$>.

<[o55-rel-o54]$_{166}$>
<[o57-rel-o56]$_{167}$>

<Mouvement [o21]$_X$> d'ouverture.

< A <l'heure [o0]$_{169}$> <où [o0]$_{170}$> <il [o1]$_{171}$> veut compter parmi les cinq leaders mondiaux de <l'assurance [o20]$_{172}$> et <où [o0]$_{173}$> <il [o1]$_{174}$> prépare < <son [o1]$_{175}$> entrée [o26]$_{177}$> à <la Bourse de Paris [o27]$_{176}$> (<le 12 juin [o6]$_{178}$> prochain), <Allianz [o1]$_{179}$> a décidé de <s' [o1]$_{180}$>essayer à la transparence.

<Il [o1]$_{181}$> publiera donc <des comptes [o49]$_X$> aux <normes [o13]$_X$> internationales IASC à <la fin [o7]$_{183}$> de <l'année [o4]$_{182}$>, bien que <celles-ci [o13]$_{184X}$> n'aient pas encore d'adaptation spécifique au <secteur [o20]$_{185}$>.

< Surtout, <le groupe [o1]$_{186}$> a, « avec deux ans d'avance sur les obligations réglementaires », publié < <ses [o1]$_{187}$> réserves [o50]$_X$> cachées, l'équivalent de <nos [o48]$_{188}$> plus-values latentes, d'un montant de <87,7 milliards [o22]$_X$> de marks (<près de 300 milliards [o22]$_{189}$> de francs). [o46]$_X$> [o21]$_{191}$>

<[o50-mde-o49]$_{190}$>

Mais les observateurs sont impatients de voir <ce mouvement [o21]$_{192}$> d'ouverture <s' [o21]$_{193}$>étendre encore dans <la comptabilité [o51]$_X$> d'<Allianz [o1]$_{194}$>, <qui [o1]$_{195}$> <se [o1]$_{196}$> refusait <hier [28 mai 1998]$_{197}$> à livrer une estimation de <sa [o1]$_{198}$> rentabilité (12,4 % en <1997 [o5]$_X$>, aux normes allemandes) fondée sur <cette nouvelle valorisation [o47]$_X$> de < <ses [o1]$_{199}$> fonds [o52]$_X$> propres.

<[o47-rel-o46]$_{201}$>
<[o51-rel-o49]$_{200}$>
<[o50-pde-o52]$_{202}$>

A.2 Annotation réponse 1

Texte B

< [vendredi 29 mai 1998] >

<La BNP [o1]$_X$> réorganise < <son [o1]$_C$> pôle [o3]$_X$> de finance-

ments spécialisés...

Comme <elle [o1]$_C$> <l' [o2]$_{ID-p}$>avait laissé entendre en <février [février 1998]$_C$>, < <la BNP [o1]$_C$> a décidé de rapprocher < <ses [o1]$_C$> filiales [o3]$_C$> de crédits spécialisés [o2]$_X$>, le Crédit Universel (biens d'équipement pour les entreprises, les professionnels et les particuliers) et BNP Bail (crédit-bail mobilier et immobilier, location).

<Le nouvel ensemble [o3]$_C$>, baptisé BNP Lease, affiche, sur la base de <1997 [année]$_\emptyset$>, un produit net bancaire de 1,7 milliard de <francs [o6]$_X$> et un résultat avant impôts de <700 millions [o6]$_C$>.

< <Sa [o3]$_C$> production [o7]$_X$> annuelle < <s' [o7]$_C$>élève [o8]$_X$> à 20 milliards de <francs [o6]$_\emptyset$> et <ses [o3]$_C$> encours < [o8]$_{ID-d}$> à <54 milliards [o6]$_C$>.

<Le nouvel ensemble [o3]$_C$>, détenu à 100 % par <la banque [o1]$_C$>, sera présidé par Claude Porcherot, tandis que Jean-René Brunon <en [o3]$_C$> assurera la direction générale.

Texte G

< [mardi 26 mai 1998, o8] >

<Guigou [o1]$_X$> visite <les locaux [o2]$_X$> du <pôle [o3]$_X$> financier

Le ban et l'arrière-ban de <la magistrature [o22]$_X$> parisienne étaient présents, <hier [lundi 25 mai 1998, o5]$_C$>, aux côtés d'<Elisabeth Guigou [o1]$_C$> pour visiter <les futurs locaux [o2]$_C$> du <pôle [o3]$_C$> financier parisien.

Après <des mois [o4]$_X$> de discussion, <le choix [o6]$_X$> définitif <s' [o6]$_C$> est porté, comme prévu, sur <le siège [o2]$_C$> historique du journal le Monde, rue des Italiens (La Tribune du <19 mai [19 mai 1998]$_C$>), au coeur du quartier parisien de <la finance [o7]$_X$>.

< [o4-rel-o5]$_{s23}$>
< [o7-rel-o3]$_{s9}$>

<Cet immeuble [o2]$_C$> luxueux, complètement réaménagé, accueillera sur <6.400 mètres [o11]$_X$> carrés, d'<ici [o8]$_C$> à <la fin [janvier 1998]$_X$> de l'année $_I$, <274 [o10]$_X$ <magistrats [o9]$_X$> et fonctionnaires >, plus <une trentaine [o12]$_X$> d'assistants spécialisés des <Finances [o7]$_{s27}$> et de la Banque de France.

< [o14-rel-o11]$_{R-ID}$>

< [{o10,o12}-mde-o13]$_C$>

<Ils [o13]$_X$> auront à <leur [o13]$_C$> disposition <quelque 23 mètres [o14]$_X$> carrés <par personne [o13,D]$_C$>, contre pratiquement <moitié [o15]$_X$> moins auparavant au Palais de justice.

< [o15-pde-o14]$_{P-R}$>

Montant du loyer : <21,6 millions [o16]$_X$> de <francs [o17]$_X$> par an <auxquels [o16]$_C$> <s' [o16]$_I$>ajoutent <15 millions [o17]$_C$> de travaux spécifiques pour sécuriser <les lieux [o2]$_C$> <que [o2]$_I$> prend en charge le propriétaire de l'immeuble, la Bred.

Outre des effectifs regroupés et supplémentaires, <le pôle [o3]$_C$> financier bénéficiera d'<équipements [o18]$_X$> informatiques.

<Les magistrats [o9]$_C$> auront notamment à <leur [o9]$_C$> disposition <des logiciels [o20]$_X$> d'instruction assistée par <ordinateur [o21]$_X$>.

< [o20-pde-o18]$_{P-M}$>
< [o21-pde-o18]$_{s10}$>
< [o9-mde-o22]$_C$>

Actuellement, <seuls [o23]$_\emptyset$> <les juges [o23]$_X$> Eva Joly et Jean-Pierre Zanotto <en [o20]$_{ID-d}$> disposent à la galerie financière de Paris.

A <ceux [o24]$_X$> <qui [o24]$_C$> <se [o24]$_C$> montrent réticents sur la mise en place de <ce pôle [o3]$_C$> financier, <le ministre [o1]$_C$> de la Justice a tenu à répéter que, « qu'on <le [o25]$_{ID-p}$> veuille ou non, <on n'échappera pas à un besoin de spécialisation croissante des <magistrats [o9]$_{ID-M}$> en matière d'information économique et financière [o25]$_X$> ».

<Cette annexe [o2]$_I$> parisienne du Palais de justice dédiée aux dossiers financiers devrait rapidement être suivie d'<autres pôles [o26]$_X$> en province.

< [o26-dde-o3]$_C$>

<Le premier [o27]$_X$> sur la liste du gouvernement est le pôle corse.

< [o27-mde-o26]$_C$>

Texte A

< [vendredi 29 mai 1998, o8] >

Fort du <rachat [o9]$_X$> des <AGF [o3]$_X$>, <Allianz [o1]$_X$> présente <son [o1]$_C$> nouveau visage

+ <L'assureur [o1]$_C$> allemand a engagé un tournant stratégique avec <le ra-

chat [o9]$_C$> des <AGF [o3]$_C$> au début de <l'année [1998]$_C$>.

+ Le développement de <sa [o1]$_C$> présence en France <l' [o1]$_C$>amènera <se [o1]$_C$> faire coter à Paris <le 12 juin [12 juin 1998]$_C$> prochain.

C'est <un nouveau groupe [o2]$_X$> Allianz <qui [o2]$_I$> naîtra d'<ici [o8]$_C$> à <la fin [janvier 1998]$_X$> de l'année $_I$.

<L'assureur [o1]$_C$> allemand, <qui [o1]$_C$> consolide <ses [o1]$_C$> 51 % des <AGF [o3]$_C$> depuis <le 1er avril [1er avril 1998]$_C$>, considère que <cette prise [o9]$_C$> de contrôle <lui [o1]$_C$> confère « <une très forte position dans le secteur de <l'assurance [o4]$_X$> mondiale, avec un pied particulièrement solide dans < <notre [o1]$_C$> marché [o5]$_X$> domestique <qu' [o5]$_C$>est l'Europe [o6]$_X$> », comme <l' [o6]$_{ID-p}$>explique <son [o1]$_C$> président, Hennig Schulte-Noelle.

< [o1-mde-o4]$_{s30}$>
< [o3-mde-o4]$_{s31}$>

Tout en <se [o1]$_C$> félicitant de < <son [o1]$_C$> acquisition [o9]$_C$>, <Allianz [o1]$_C$> n'en rappelle pas moins <ses [o1]$_C$> objectifs.

Contre-pied d'<Axa [o7]$_X$>.

< [o7-mde-o4]$_{s32}$>

D'<ici [o8]$_C$> à <2000 [année, o25]$_\emptyset$>, le résultat net des <AGF [o3]$_C$> devra être porté à 5,5 milliards de francs, permettant à <sa [o3]$_C$> maison mère de constater un retour sur investissement de 9 %.

Pour <1998 [o25]$_\emptyset$>, <l'acquisition [o9]$_C$> des <AGF [o3]$_C$> doit permettre à <Allianz [o1]$_C$> de porter <son [o1]$_C$> chiffre d'affaires à <107 milliards [o10]$_X$> de deutsche marks (<358 milliards [o11]$_X$> de francs) et de constater une croissance à deux chiffres de <son [o1]$_C$> bénéfice net.

< [o10-rel-o11]$_{R-ID}$>

Mais pour <l'heure [o8]$_C$>, la priorité est à <la « digestion » de <cette opération [o9]$_C$> d'<ici [o8]$_C$> à <fin 1999 [o25]$_\emptyset$> [o12]$_X$> — <ce [o12]$_C$> <qui [o12]$_C$> passe par la détermination, <cette année [1998]$_C$>, des structures, des cadres dirigeants et des plans de développement dans chaque pays.

Au passage, <l'assureur [o1]$_C$> allemand rappelle <sa [o1]$_C$> volonté de décentralisation et, comme pour prendre le contre-pied de <son [o1]$_C$> grand concurrent <Axa [o7]$_C$>, affirme qu'<il [o1]$_C$> ne voit pas l'intérêt de coiffer <ses [o1]$_C$> sociétés nationales d'une marque mondiale.

A <l'heure [o8]$_C$> actuelle, aucune décision n'a été prise quant à d'éventuelles cessions d'actifs, et notamment <les 25 % [cession,o13]$_X$> de la Coface,

comme <le [o13]$_{ID-p}$> demande Bruxelles.

Face à cette nouvelle donne, <Allianz [o1]$_C$> réfléchit à la façon de parvenir à une présentation commune et aux économies possibles entre Euler et Hermès, sans pour autant créer une structure faîtière.

Enfin, toujours en France, <Hennig Schulte-Noelle [o14]$_X$> a réaffirmé <son [o14]$_C$> intention d'entrer jusqu'à 10 % dans le capital du Crédit Lyonnais.

D'un point de vue stratégique, <cette année [1998]$_C$> devrait marquer, selon <Allianz [o1]$_C$>, le démarrage de <sa [o1]$_C$> coopération dans la gestion d'actifs avec <la Dresdner Bank [o15]$_X$>, pour <laquelle [o15]$_\emptyset$> la répartition future des compétences reste en négociation.

En assurance, <Allianz [o1]$_C$> discute avec la nouvelle Bayerische und Vereinsbank <une répartition [o16]$_X$> des partenariats entre <lui [o1]$_C$> et <Ergo [o18]$_X$>.

A <ce sujet [o16]$_C$>, et pour tenter de briser le soupçon de cartel porté sur le marché allemand, <Allianz [o1]$_C$> rappelle que <ses [o1]$_C$> 10 % d'<Ergo [o18]$_C$> ne sont pas stratégiques et que <les deux compagnies [o17]$_X$> restent « des concurrentes acharnées ».

< [{o1,o18}-mde-o17]$_C$>

Du reste, <Allianz [o1]$_C$> (<qui [o1]$_C$> n'a connu en <1997 [o25]$_\emptyset$> qu'une hausse de <ses [o1]$_C$> primes de 0,4 % sur le marché automobile allemand en croissance moyenne de 1,3 %), affirme qu'<il [o1]$_C$> « continuera à prendre des mesures pour maintenir, sinon augmenter, <sa [o1]$_C$> part de marché sur ce segment très concurrentiel ».

Autant dire que la guerre des tarifs en Allemagne n'est pas terminée.

En <1997 [o25]$_\emptyset$>, <l'assureur [o1]$_C$>, <qui [o1]$_C$> a publié un bénéfice définitif de 2,7 milliards de <marks [o19]$_X$> (+ 20,5 %), a, pour la première fois, affiché <l'an [1997]$_C$> dernier un bénéfice technique positif de <182 millions [o19]$_C$>.

Mais en assurance-dommages, le chiffre reste négatif de <1,3 milliard [o19]$_C$>.

<Mouvement [o26]$_X$> d'ouverture.

A <l'heure [o8]$_C$> où <il [o1]$_C$> veut compter parmi <les cinq leaders [o20]$_X$> mondiaux de <l'assurance [o4]$_C$> et où <il [o1]$_C$> prépare <son [o1]$_C$> entrée à la Bourse de Paris (<le 12 juin [12 juin 1998]$_C$> prochain), <Allianz [o1]$_C$> a décidé de <s' [o1]$_C$>essayer à la transparence.

< [o20-mde-o4]$_T$>

<Il [o1]$_C$> publiera donc des comptes aux normes internationales IASC à <la fin [janvier 1998]$_I$> de l'année $_I$, bien que <celles-ci [normes]$_{d-ID}$> n'aient

pas encore d' adaptation spécifique au secteur.

Surtout, <le groupe [o1]$_C$> a, « avec <deux ans [2000]$_\emptyset$> d'avance sur les obligations réglementaires », publié <ses [o1]$_C$> réserves cachées, l'équivalent de nos plus-values latentes, d'un montant de <87,7 milliards [o21]$_X$> de marks (près de <300 milliards [o22]$_X$> de francs).

< [o21-rel-o22]$_{R-ID}$>

Mais les observateurs sont impatients de voir <ce mouvement [o26]$_C$> d'ouverture <s' [o26]$_C$>étendre encore dans la comptabilité d'<Allianz [o1]$_C$>, <qui [o1]$_C$> <se [o1]$_C$> refusait <hier [o23]$_X$> à livrer une estimation de <sa [o1]$_C$> rentabilité (12,4 % en <1997 [o25]$_\emptyset$>, aux normes allemandes) fondée sur <cette nouvelle valorisation [o24]$_X$> de <ses [o1]$_C$> fonds propres.

< [o24-pde-o26]$_{s33}$>
< [o23-dde-o8]$_{s29}$>

A.3 Annotation réponse 2

Texte B

< [vendredi 29 mai 1998] >

<La BNP [o1]$_X$> réorganise < <son [o1]$_C$> pôle [o2]$_X$> de <financements [o3]$_X$> spécialisés...

Comme <elle [o1]$_C$> <l' [o10(P)]$_C$>avait laissé entendre en <février [fevrier 1998]$_C$>, < <la BNP [o1]$_C$> a décidé de rapprocher < <ses [o1]$_C$> filiales [o2]$_C$> de <crédits [o3]$_C$> spécialisés [o10]$_X$>, <le Crédit Universel [o4]$_X$> (<biens [o5]$_X$> d'équipement pour les entreprises, les professionnels et les particuliers) et <BNP Bail [o6]$_X$> (<crédit-bail [o7]$_X$> mobilier et immobilier, <location [o8]$_X$>).

< [{o4,o6}-mde-o2]$_T$>
< [o5-rel-o3]$_{R-M}$>
< [o7-rel-o3]$_{R-M}$>
< [o8-rel-o3]$_{R-M}$>

<Le nouvel ensemble [o2]$_C$>, baptisé BNP Lease, affiche, sur la base de 1997, <un produit [o13]$_X$> net bancaire de 1,7 milliard de francs et <un résultat [o14a]$_X$> avant impôts de 700 millions.

< [o14a-pde-o13]$_{s2}$>

< <Sa [o1]$_I$> production [o9]$_X$> annuelle < <s' [o9]$_C$> élève [s'élever]$_X$ à 20 milliards de francs et < <ses [o1]$_I$> encours [o14b]$_X$> < [s' élever]$_C$>

à 54 milliards.

< [o13-pde-o9]$_{s3}$>
< [o14b-rel-o9]$_{s4}$>

<Le nouvel ensemble [o2]$_C$>, détenu à 100 % par <la banque [o1]$_C$>, sera présidé par <Claude Porcherot [o11]$_X$>, tandis que <Jean-René Brunon [o12]$_X$> <en [o2]$_C$> assurera la direction générale.

< [o11-pde-o2]$_T$>
< [o12-pde-o2]$_T$>
< [o2-pde-o1]$_T$>

Texte G

< [mardi 26 mai 1998] >

<Guigou [o1]$_X$> visite <les locaux [o2]$_X$> du <pôle [o3]$_X$> financier

<Le ban [o4]$_X$> et <l'arrière-ban [o5]$_X$> de <la magistrature [o6]$_X$> parisienne étaient présents, <hier [lundi 25 mai 1998]$_C$>, aux côtés d'<Elisabeth Guigou [o1]$_C$> pour visiter <les futurs locaux [o2]$_C$> du <pôle [o3]$_C$> financier parisien.

< [{o4,o5}-mde-o6]$_T$>

Après des mois de discussion, <le choix [o7]$_X$> définitif <s' [o7]$_C$> est porté, comme prévu, sur <le siège [o2]$_C$> historique du journal le Monde, <rue [o8]$_X$> des Italiens (La Tribune du <19 mai [mardi 19 mai 1998, o11]$_C$>), au coeur du <quartier [o9]$_X$> parisien de la finance.

< [o8-pde-o9]$_T$>

<Cet immeuble [o2]$_C$> luxueux, complètement réaménagé, accueillera sur <6.400 mètres [o29, espace]$_X$> carrés, d'<ici [o11]$_I$> à <la fin [decembre 1998]$_X$> de l'année $_C$, 274 <magistrats [o12]$_X$> et <fonctionnaires [o13]$_X$> , plus <une trentaine d'assistants [o14]$_X$> spécialisés des Finances et de la Banque de France.

< [o12-mde-o6]$_C$>

<Ils [o15]$_X$> auront à <leur [o15]$_C$> disposition <quelque 23 mètres [o28]$_X$> carrés <par personne [o15(D)]$_C$>, contre pratiquement <moitié [espace]$_C$> moins auparavant au <Palais [o24]$_X$> de justice.

< [{o12,o13,o14}-mde-o15]$_C$>
< [o28-pde-o29]$_{P-ID}$>

Montant du loyer : <21,6 millions [o25]$_X$> de francs par an <auxquels [o25]$_C$>

<s' [o16]$_C$>ajoutent <15 millions [o16]$_X$> de <travaux [o10]$_X$> spécifiques pour sécuriser <les lieux [o2]$_C$> <que [o10]$_C$> prend en charge le propriétaire de <l'immeuble [o2]$_C$>, la Bred.

Outre <des effectifs [o15]$_C$> regroupés et supplémentaires, <le pôle [o3]$_C$> financier bénéficiera d'<équipements [o17]$_X$> informatiques.

<Les magistrats [o12]$_C$> auront notamment à <leur [o12]$_C$> disposition <des logiciels [o17]$_{ID-M}$> d'instruction assistée par <ordinateur [o18G]$_X$>.

< [o17-mde-o18]$_{s11}$>

Actuellement, seuls <les juges [o30]$_X$> Eva Joly et Jean-Pierre Zanotto <en [o17]$_{ID-d}$> disposent à <la galerie [o23]$_X$> financière de Paris.

< [o23-pde-o24]$_{s12}$>
< [o30-pde-o6]$_{P-M}$>

A <ceux [o19]$_X$> qui <se [o19]$_C$> montrent réticents sur la mise en place de <ce pôle [o3]$_C$> financier, <le ministre [o1]$_C$> de la Justice a tenu à répéter que, « qu'<on [o20]$_X$> <le [o21]$_{ID-p}$> veuille ou non, <on [o20]$_C$> n'échappera pas à un besoin de <spécialisation [o21]$_X$> croissante des <magistrats [o22G]$_X$> en matière d'information économique et financière ».

< [o19-pde-o15]$_{s13}$>
< [o15-pde-o20]$_{s14}$>
< [o12-pde-o22]$_{P-M}$>

<Cette annexe [o2]$_I$> parisienne du <Palais [o3]$_I$> de justice dédiée aux dossiers financiers devrait rapidement être suivie d'<autres pôles [o25]$_X$> en province.

< [o25-dde-o3]$_C$>

<Le premier [o26]$_X$> sur la liste du <gouvernement [o27]$_X$> est le pôle corse.

< [o26-mde-o25]$_C$>
< [o1-mde-o27]$_C$>

Texte A

< [vendredi 29 mai 1998] >

Fort du <rachat [o1]$_X$> des <AGF [o2]$_X$>, <Allianz [o3]$_X$> présente < <son [o4]$_X$> nouveau visage [o5]$_X$>

+ <L'assureur [o3]$_C$> allemand a engagé un tournant stratégique avec <le rachat [o1]$_C$> des <AGF [o2]$_C$> au <début [janvier 1998]$_T$> de <l'année

[1998]$_C$>.

+ Le développement de <sa [o3]$_C$> présence en France <l' [o3]$_C$> amènera <se [o3]$_C$> faire coter à Paris <le 12 juin [12 juin 1998]$_C$> prochain.

C'est <un nouveau groupe [o5]$_X$> Allianz <qui [o5]$_I$> naîtra d'<ici [12 juin 1998]$_I$> à <la fin [decembre 1998]$_X$> de l'année $_C$.

<L'assureur [o3]$_C$> allemand, <qui [o3]$_C$> consolide <ses [o3]$_C$> 51 % des <AGF [o2]$_C$> depuis <le 1er avril [1er avril 1998]$_C$>, considère que <cette prise [o1]$_C$> de contrôle <lui [o3]$_C$> confère « <une très forte position [o6]$_X$> dans le secteur de <l'assurance [o22]$_X$> mondiale, avec un pied particulièrement solide dans <notre [o3]$_C$> marché domestique <qu' [o29]$_C$>est <l'Europe [o29]$_X$> », comme <l' [o6]$_{ID-p}$>explique < <son [o3]$_C$> président [o10]$_X$>, Hennig Schulte-Noelle.

Tout en <se [o3]$_C$> félicitant de < <son [o3]$_C$> acquisition [o22]$_I$>, <Allianz [o3]$_C$> n'<en [o22]$_{s51}$> rappelle pas moins < <ses [o3]$_C$> objectifs [o19]$_X$>.

< [{o17,o16,o23,o24}-mde-o19]$_{s34,s35,C,C}$>

<Contre-pied [o17]$_X$> d'<Axa [o8]$_X$>.

D'<ici [1998]$_I$> à 2000, <le résultat [o20]$_X$> net des <AGF [o2]$_C$> devra être porté à 5,5 milliards de francs, permettant à < <sa [o2]$_C$> maison [o3]$_C$> mère de constater <un retour [o21]$_X$> sur investissement de 9 %.

< [o21-pde-o20]$_{s36}$> [1]

Pour <1998 [1998]$_T$>, <l'acquisition [o1]$_C$> des <AGF [o2]$_C$> doit permettre à <Allianz [o3]$_C$> de porter < <son [o3]$_C$> chiffre [o22b]$_X$> d'affaires à 107 milliards de deutsche marks (358 milliards de francs) et de constater une croissance à deux chiffres de < <son [o3]$_C$> bénéfice [o20]$_{s53}$> net.

< [o20-pde-o22b]$_{s37}$> [2]

Mais pour <l'heure [mai 1998]$_I$>, la priorité est à <la « digestion » [o7]$_X$> de <cette opération [o1]$_C$> d'<ici [12 juin 1998]$_I$> à <fin 1999 [decembre 1999]$_T$> — <ce [o7]$_C$> qui passe par la détermination, <cette année [1998]$_C$>, des structures, des cadres dirigeants et des plans de développement dans chaque pays.

Au passage, <l'assureur [o3]$_C$> allemand rappelle <sa [o3]$_C$> volonté de <décentralisation [o24]$_X$> et, comme pour prendre <le contre-pied [o17]$_C$> de < <son [o3]$_C$> grand concurrent [o8]$_C$> Axa, affirme qu'<il [o3]$_C$> ne voit pas

[1] Note de l'annotateur : retour sur investissement = résultat net - inverstissement
[2] Note de l'annotateur : benefice = chiffre d'affaires - frais

l'intérêt de coiffer <ses [o3]$_C$> sociétés nationales d'une marque mondiale.

A <l'heure [mai 1998]$_I$> actuelle, <aucune décision [o9(D)]$_{s50}$> n'a été prise quant à <d'éventuelles cessions [o9]$_X$> d'actifs, et notamment les 25 % de la Coface, comme <le [o9]$_{ID-p}$> demande Bruxelles.

Face à <cette nouvelle donne [o24]$_{s54}$>, <Allianz [o3]$_C$> réfléchit à la façon de parvenir à une présentation commune et aux économies possibles entre Euler et Hermès, sans pour autant créer une structure faîtière.

Enfin, toujours en France, <Hennig Schulte-Noelle [o10]$_C$> a réaffirmé < <son [o10]$_C$> intention [o23]$_X$> d'entrer jusqu'à 10 % dans le capital du Crédit Lyonnais.

D'un point de vue stratégique, <cette année [1998]$_C$> devrait marquer, selon <Allianz [o3]$_C$>, le démarrage de <sa [o3]$_C$> coopération dans la gestion d'actifs avec <la Dresdner Bank [o25]$_X$>, pour <laquelle [o25]$_\emptyset$> la répartition future des compétences reste en négociation.

En <assurance [o28(G)]$_{s52}$>, <Allianz [o3]$_C$> discute avec la nouvelle Bayerische und Vereinsbank <une répartition [o11]$_X$> des partenariats entre <lui [o3]$_C$> et <Ergo [o12]$_X$>.

A <ce sujet [o11]$_C$>, et pour tenter de briser le soupçon de cartel porté sur le marché allemand, <Allianz [o3]$_C$> rappelle que <ses [o3]$_C$> 10 % d'<Ergo [o12]$_C$> ne sont pas stratégiques et que <les deux compagnies [o13]$_X$> restent « des concurrentes acharnées ».

< [{o3,o12}-mde-o13]$_C$>

Du reste, <Allianz [o3]$_C$> (<qui [o3]$_C$> n'a connu en 1997 qu'une hausse de <ses [o3]$_C$> primes de 0,4 % sur <le marché [o26]$_X$> automobile allemand en croissance moyenne de 1,3 %), affirme qu'<il [o3]$_C$> « continuera à prendre des mesures pour maintenir, sinon augmenter, <sa [o3]$_C$> part de marché sur <ce segment [o26]$_C$> très concurrentiel ».

Autant dire que la guerre des tarifs en Allemagne n'est pas terminée.

En 1997, <l'assureur [o3]$_C$>, <qui [o3]$_C$> a publié <un bénéfice [o27]$_X$> définitif de 2,7 milliards de marks (+ 20,5 %), a, pour la première fois, affiché <l'an [1997]$_C$> dernier <un bénéfice [o14]$_X$> technique positif de 182 millions.

< [o27-dde-o14]$_{s43}$>

Mais en assurance-dommages, <le chiffre [o14]$_{ID-R}$> reste négatif de 1,3 milliard.

<Mouvement [o16]$_X$> d'ouverture.

A <l'heure [mai 1998]$_I$> où <il [o3]$_C$> veut compter parmi les cinq leaders mondiaux de <l'assurance [o28]$_X$> et <où [mai 1998]$_I$> <il [o3]$_C$> prépare

<son [o3]$_C$> entrée à la Bourse de Paris (<le 12 juin [12 juin 1998]$_C$> prochain), <Allianz [o3]$_C$> a décidé de <s' [o3]$_C$>essayer à la transparence.

<Il [o3]$_C$> publiera donc des comptes aux <normes [o15]$_X$> internationales IASC à <la fin [decembre 1998]$_C$> de l'année $_C$, bien que <celles-ci [o15]$_C$> n'aient pas encore d' adaptation spécifique au <secteur [o28]$_C$>.

Surtout, <le groupe [o3]$_C$> a, « avec deux ans d'avance sur les obligations réglementaires », publié <ses [o3]$_C$> réserves cachées, l'équivalent de <nos [France]$_C$> plus-values latentes, d'un montant de 87,7 milliards de marks (près de 300 milliards de francs).

Mais les observateurs sont impatients de voir <ce mouvement [o16]$_C$> d'ouverture <s' [o16]$_C$>étendre encore dans la comptabilité d'<Allianz [o3]$_C$>, qui <se [o3]$_C$> refusait <hier [jeudi 28 mai]$_C$> à livrer une estimation de < <sa [o3]$_C$> rentabilité [o18]$_X$> (12,4 % en 1997, aux normes allemandes) fondée sur cette nouvelle valorisation de <ses [o3]$_C$> fonds propres.

< [o18-rel-o22b]$_{s38}$> [3]

A.4 Annotation réponse 3

Texte B

< [vendredi 29 mai 1998] >

<La BNP [o1]$_X$> réorganise <son pôle [o2]$_X$> de financements spécialisés...

Comme <elle [o1]$_C$> <l' [o6]$_{ID-p}$>avait laissé entendre en <février [fevrier 1998]$_C$>, la BNP a décidé de <rapprocher <ses filiales [o5]$_X$> [o6]$_X$> de crédits spécialisés, le Crédit Universel(biens d'équipement pour les entreprises , les professionnels et les particuliers) et BNP Bail (crédit-bail mobilier et immobilier, location).

<Le nouvel ensemble [o4]$_X$>, baptisé BNP Lease, affiche, sur la base de 1997, un produit net bancaire de 1,7 milliard de francs et un résultat avant impôts de 700 millions.

< [{o5,o1}-mde-o4]$_{M-ID,s1}$>

<Sa [o4]$_C$> production annuelle s'élève à 20 milliards de francs et <ses [o4]$_C$> encours à 54 milliards.

<Le nouvel ensemble [o4]$_C$>, détenu à 100 % par <la banque [o1]$_C$>, sera pré-

[3] Note de l'annotateur : rentabilité = %(gains/chiffre d'affaire)

sidé par Claude Porcherot, tandis que Jean-René Brunon <en [o4]$_C$> assurera <la direction [o7]$_X$> générale.

< [o7-pde-o4]$_T$>

Texte G

< [mardi 26 mai 1998] >

Guigou visite <les locaux [o1]$_X$> du <pôle [o2]$_X$> financier

Le ban et l'arrière-ban de la magistrature parisienne étaient présents, <hier [25 mai 1998]$_C$>, aux côtés d'Elisabeth Guigou pour visiter <les futurs locaux [o1]$_C$> du pôle financier parisien.

Après des mois de discussion, le choix définitif s'est porté, comme prévu, sur <le siège [o1]$_C$> historique du journal le Monde, rue des Italiens (La Tribune du 19 mai), au coeur du quartier parisien de la finance.

<Cet immeuble [o1]$_C$> luxueux, complètement réaménagé, accueillera sur <6.400 mètres [o9]$_X$> carrés, d'<ici [mardi 26 mai 1998]$_C$> à la fin de <l'année [decembre 1998]$_C$>, <274 <magistrats [o4]$_X$> et <fonctionnaires [o5]$_X$> , plus <une trentaine [o6]$_X$> d'assistants spécialisés des Finances et de la Banque de France [o7]$_X$>.

< [o4-mde-o2]$_T$>
< [o5-mde-o2]$_T$>
< [o6-mde-o2]$_T$>

<Ils [o7]$_C$> auront à <leur [o7]$_C$> disposition <quelque 23 mètres [o8]$_X$> carrés <par personne [o10]$_X$>, contre pratiquement moitié moins auparavant au Palais de justice .

< [o8-mde-o9]$_{M-ID}$>
< [o10-mde-o7]$_{M-ID}$>

Montant du loyer : <21,6 millions [o11]$_X$> de francs par an <auxquels [o11]$_C$> s' ajoutent <15 millions [o12]$_X$> de travaux spécifiques pour sécuriser <les lieux [o1]$_C$> <que [o12]$_C$> prend en charge le propriétaire de <l'immeuble [o1]$_C$>, la Bred.

Outre des effectifs regroupés et supplémentaires, <le pôle [o2]$_C$> financier bénéficiera d'<équipements [o13]$_X$> informatiques.

<Les magistrats [o4]$_C$> auront notamment à <leur [o4]$_C$> disposition <des

logiciels [o14]$_X$> d'instruction assistée par ordinateur .

< [o14-mde-o13]$_C$>

<Actuellement [mai 1998]$_I$>, seuls les juges Eva Joly et Jean-Pierre Zanotto <en [o14]$_{ID-d}$> disposent à la galerie financière de Paris.

A ceux qui se montrent réticents sur la mise en place de <ce pôle [o2]$_C$> financier, le ministre de la Justice a tenu à répéter que, « qu'on <le [o15]$_{ID-p}$> veuille ou non,< on n'échappera pas à un besoin de <spécialisation [o2]$_{s24}$> croissante des magistrats en matière d'information économique et financière » [o15]$_X$>.

<Cette annexe [o1]$_I$> parisienne du Palais de justice dédiée aux dossiers financiers devrait rapidement être suivie d'<autres pôles [o16]$_X$> en province.

< [o16-mde-o2]$_{M-D}$>

<Le premier [o17]$_X$> sur la liste du gouvernement est le pôle corse.

< [o17-mde-o16]$_C$>

Texte A

< [vendredi 29 mai 1998] >

Fort du <rachat [o6]$_X$> des AGF, <Allianz [o1]$_X$> présente <son [o1]$_I$> nouveau visage

+ <L'assureur [o2]$_X$> allemand a engagé un tournant stratégique avec le rachat des AGF au début de <l'année [1998]$_C$>.

+ Le développement de <sa [o2]$_I$> présence en France <l' [o2]$_I$>amènera <se [o2]$_I$> faire coter à Paris <le 12 juin [12 juin 1998]$_C$> prochain.

C'est <un nouveau groupe [o3]$_X$> Allianz <qui [o3]$_C$> naîtra d'ici à <la fin de l'année [31 decembre 1998]$_{X,C}$>.

<L'assureur [o2]$_I$> allemand, <qui [o2]$_I$> consolide < <ses [o2]$_I$> 51 % [o3]$_{s41}$> des AGF depuis <le 1er avril [1 avril 1998]$_C$>, considère que <cette prise [o3]$_I$> de contrôle <lui [o2]$_I$> <confère « une très forte position [o5]$_X$> dans le secteur de l'assurance mondiale, avec un pied particulièrement solide dans <notre marché [o4]$_{s55}$> domestique <qu' [o4]$_I$>est l'Europe », comme <l' [o5]$_{ID-p}$>explique <son [o2]$_I$> président, Hennig Schulte-Noelle.

Tout en <se [o3]$_C$> félicitant de < <son [o3]$_C$> acquisition [o4]$_I$>, <Allianz

[o3]$_C$> n'<en [o4]$_{s51}$> rappelle pas moins < <ses [o3]$_C$> objectifs [o6]$_{s57}$>.

Contre-pied d'Axa.

D'<ici [vendredi 29 mai 1998]$_C$> à 2000, le résultat net des <AGF [o4]$_X$> devra être porté à <5,5 milliards [o7]$_X$> de francs , permettant à <sa [o4]$_C$> maison mère de constater un retour sur investissement de <9 % [o8]$_X$>.

< [o8-rel-o7]$_{s39}$>

Pour 1998, <l'acquisition [o6]$_C$> des AGF doit permettre à <Allianz [o3]$_C$> de porter <son [o3]$_C$> chiffre d'affaires à 107 milliards de deutsche marks (358 milliards de francs) et de constater une croissance à deux chiffres de <son [o3]$_C$> bénéfice net.

Mais pour <l'heure [29 mai 1998]$_C$>, la priorité est à <la « digestion » [o9]$_X$> de <cette opération [o6]$_C$> d'<ici [29 mai 1998]$_C$> à <fin 1999 [31 decembre 1999]$_T$> — <ce [o9]$_C$> qui passe par <la détermination [o10]$_X$>, <cette année [1998]$_C$>, des structures, des cadres dirigeants et des plans de développement dans chaque pays.

< [o10-pde-o9]$_T$>

Au passage, <l'assureur [o2]$_I$> allemand rappelle <sa [o2]$_I$> volonté de décentralisation et, comme pour prendre le contre-pied de <son [o2]$_I$> grand concurrent Axa, affirme qu'<il [o2]$_I$> ne voit pas l'intérêt de coiffer <ses [o2]$_I$> sociétés nationales d'une marque mondiale.

A <l'heure [29 mai 1998]$_C$> actuelle, <aucune décision [o13]$_X$> n'a été prise quant à <d'éventuelles cessions [o11]$_X$> d'actifs, et notamment <les 25 % [o12]$_X$> de la Coface, comme <le [o11]$_{ID-p}$> demande Bruxelles.

< [o12-mde-o11]$_C$>

Face à <cette nouvelle donne [o13]$_{s54}$>, Allianz réfléchit à la façon de parvenir à une présentation commune et aux économies possibles entre Euler et Hermès, sans pour autant créer une structure faîtière.

Enfin, toujours en France, <Hennig Schulte-Noelle [o14]$_X$> a réaffirmé <son [o14]$_C$> intention d'entrer jusqu'à 10 % dans le capital du Crédit Lyonnais.

D'un point de vue stratégique, <cette année [1998]$_C$> devrait marquer, selon Allianz, le démarrage de <sa [o3]$_C$> coopération dans la gestion d'actifs avec <la Dresdner Bank [o15]$_X$>, pour <laquelle [o15]$_\emptyset$> la répartition future des compétences reste en négociation.

En assurance, Allianz discute avec <la nouvelle Bayerische und Vereinsbank [o18]$_X$> <une répartition [o16]$_X$> des partenariats entre <lui [o3]$_C$> et Ergo.

A <ce sujet [o16]$_C$>, et pour tenter de briser le soupçon de cartel porté sur <le marché [o21]$_X$> allemand, Allianz rappelle que <ses [o3]$_C$> 10 % d'<Ergo

[o17]$_X$> ne sont pas stratégiques et que <les deux compagnies [o19]$_X$> restent « des concurrentes acharnées ».

< [{o17,o18}-mde-o19]$_I$>

Du <reste [o20]$_X$>, Allianz (<qui [o3]$_C$> n'a connu en 1997 qu'une hausse de < <ses [o3]$_C$> primes [o23]$_X$> de 0,4 % sur <le marché [o22]$_X$> automobile allemand en croissance moyenne de 1,3 %), affirme qu'<il [o3]$_C$> « continuera à prendre des mesures pour maintenir, sinon augmenter, < <sa [o3]$_C$> part [o23]$_{s56}$> de marché sur <ce segment [o22]$_C$> très concurrentiel ».

< [o20-dde-o16]$_{s49}$>

< [o22-mde-o21]$_C$>

Autant dire que la guerre des tarifs en Allemagne n'est pas terminée.

En 1997, <l'assureur [o24]$_X$>, <qui [o24]$_I$> a publié un bénéfice définitif de 2,7 milliards de marks (+ 20,5 %), a, pour la première fois, affiché <l'an [1997]$_C$> dernier un bénéfice technique positif de 182 millions.

Mais en assurance-dommages, le chiffre reste négatif de 1,3 milliard.

Mouvement d'ouverture.

A <l'heure [o25]$_X$> <où [o25]$_I$> <il [o3]$_C$> veut compter parmi les cinq leaders mondiaux de <l'assurance [o27]$_X$> et <où [o25]$_I$> <il [o3]$_C$> prépare <son [o3]$_C$> entrée à la Bourse de Paris (<le 12 juin [12 juin 1998]$_C$> prochain), Allianz a décidé de <s' [o3]$_C$>essayer à la transparence.

<Il [o3]$_C$> publiera donc des comptes aux <normes [o26]$_X$> internationales IASC à la fin de <l'année [31 decembre 1998]$_{C,C}$>, bien que <celles-ci [o26]$_C$> n'aient pas encore d' adaptation spécifique au <secteur [o27]$_C$>.

Surtout,< <le groupe [o3]$_C$> a, « avec deux ans d'avance sur les obligations réglementaires », publié ses réserves cachées, l'équivalent de nos plus-values latentes, d'un montant de 87,7 milliards de marks (près de 300 milliards de francs). [o28]$_C$>

Mais les observateurs sont impatients de voir <ce mouvement [o28]$_X$> d'ouverture <s' [o28]$_C$>étendre encore dans <la comptabilité [o29]$_X$> d'Allianz, <qui [o3]$_C$> <se [o3]$_C$> refusait hier à livrer une estimation de < <sa [o3]$_C$> rentabilité [o30]$_X$> (12,4 % en 1997, aux normes allemandes) fondée sur cette nouvelle valorisation de ses fonds propres.

< [o29-pde-o3]$_T$>

< [o30-rel-o29]$_{s40}$>

A.5 Annotation réponse 4

Texte B

< [vendredi 29 mai 1998] >

<La BNP [o1]$_X$> réorganise < <son [o1]$_C$> pôle [o2]$_X$> de financements spécialisés...

Comme <elle [o1]$_C$> <l' [o1]$_I$>avait laissé entendre en <février [fevrier 1998]$_C$>, <la BNP [o1]$_C$> a décidé de rapprocher < <ses [o1]$_C$> filiales [o2]$_C$> de crédits spécialisés, le Crédit Universel (biens d'équipement pour les entreprises, les professionnels et les particuliers [o3]$_X$) et BNP Bail (crédit-bail mobilier et immobilier, location [o4]$_X$).

< [{o3,o4}-mde-o2]$_\emptyset$>
< [o2-pde-o1]$_T$>

<Le nouvel ensemble [o2]$_C$>, baptisé BNP Lease, affiche, sur <la base [o10]$_X$> de <1997 [1997]$_T$>, <un produit [o8]$_X$> net bancaire de 1,7 milliard de francs et <un résultat [o9]$_X$> avant impôts de 700 millions.

< [o8-rel-o10]$_{s5}$>
< [o9-rel-o10]$_{s6}$>
< [o10-rel-o2]$_{s7}$> [4]

< <Sa [o2]$_C$> production [o6]$_X$> annuelle <s' [o6]$_C$>élève à 20 milliards de francs et < <ses [o2]$_C$> encours [o7]$_X$> à 54 milliards.

< [o6-pde-o2]$_T$>
< [o7-pde-o2]$_T$>

<Le nouvel ensemble [o2]$_C$>, détenu à 100 % par <la banque [o1]$_C$>, sera présidé par Claude Porcherot, tandis que Jean-René Brunon en assurera <la direction [o5]$_X$> générale.

< [o5-pde-o2]$_T$>

Texte G

< [mardi 26 mai 1998]>

Guigou visite <les locaux [o1]$_X$> du <pôle [o4]$_X$> financier

<Le ban [o11]$_X$> et <l'arrière-ban [o12]$_X$> de <la magistrature [o13]$_X$> pa-

[4] Le produit et le résultat sont calculés à partir de l'exercice 97 et des prévisions envisagées par le rapprochement de ces deux filiales.

risienne étaient présents, <hier [lundi 25 mai 1998]$_C$>, aux <côtés [o13]$_{s26}$> d'<Elisabeth Guigou [o14]$_X$> pour visiter les futurs <locaux [o1]$_C$> du <pôle [o4]$_C$> financier parisien.

< [o11-pde-o13]$_T$>

< [o12-pde-o13]$_T$>

< [o11-dde-o12]$_T$>

Après des mois de discussion, <le choix [o14]$_X$> définitif <s' [o14]$_C$>est porté, comme prévu, sur <le siège [o1]$_C$> historique du journal le Monde, rue des Italiens (La Tribune du <19 mai [mardi 19 mai 1998]$_C$>), au coeur du quartier parisien de la finance .

<Cet immeuble [o1]$_C$> luxueux, complètement réaménagé, accueillera sur 6.400 mètres carrés, d'<ici [lundi 25 mai 1998]$_I$> à <la fin de l'année [decembre 1998]$_C$>, <274 <magistrats [o7]$_X$> et <fonctionnaires [o8]$_X$> [o2]$_X$>, plus <une trentaine d'assistants [o3]$_X$> spécialisés des Finances et de la Banque de France .

< [{o2,o3}-mde-o4]$_{s25}$>

< [{o7,o8}-mde-o2]$_T$>

<Ils [o4]$_I$> auront à <leur [o4(D)]$_I$> disposition <quelque 23 mètres [surface]> carrés par <personne [o4(D)]$_I$>, contre pratiquement <moitié [surface]$_C$> moins auparavant au Palais de justice.

<Montant [o6]$_X$> du <loyer [o1]$_{ID-R}$> : 21,6 millions de <francs [francs]$_X$> par an <auxquels [o6]$_C$> <s' [o6]$_I$>ajoutent <15 millions [francs]$_C$> de travaux spécifiques pour sécuriser <les lieux [o1]$_C$> que prend en charge le propriétaire de <l'immeuble [o1]$_C$>, la Bred.

Outre <des effectifs [o4]$_I$> regroupés et supplémentaires, <le pôle [o4]$_C$> financier bénéficiera d'<équipements [o9]$_X$> informatiques.

<Les magistrats [o7]$_C$> auront notamment à <leur [o7(D)]$_C$> disposition <des logiciels [o10]$_X$> d'instruction assistée par ordinateur.

< [o10-pde-o9]$_{P-M}$>

Actuellement, seuls les juges Eva Joly et Jean-Pierre Zanotto <en [o10]$_{ID-d}$> disposent à la galerie financière de Paris.

A <ceux [o21]$_X$> qui <se [o21]$_C$> montrent réticents sur la mise en place de <ce pôle [o4]$_C$> financier, <le ministre [o14]$_C$> de la Justice a tenu à répéter que, « qu'<on [o19]$_X$> <le [o20]$_{ID-p}$> veuille ou non, <on [o19]$_C$> n'échappera pas à <un besoin [o20]$_X$> de spécialisation croissante des <magistrats

[o15(G)]$_X$> en matière d'information économique et financière ».

< [o7-mde-o15]$_C$>

<Cette annexe [o1]$_I$> parisienne du Palais de justice dédiée aux dossiers financiers devrait rapidement être suivie d'<autres pôles [o16]$_X$> en province.

< [o4-dde-o16]$_C$>

<Le premier [o17]$_X$> sur <la liste [o16]$_{ID-R}$> du <gouvernement [o18]$_X$> est le pôle corse.

< [o17-mde-o16]$_C$>
< [o14-mde-o18]$_C$>

Texte A

< [vendredi 29 mai 1998, o10] >

Fort du <rachat [o2]$_X$> des AGF, <Allianz [o1]$_X$> présente <son [o1]$_C$> nouveau visage

+ <L'assureur [o1]$_C$> allemand a engagé un tournant stratégique avec <le rachat [o2]$_C$> des AGF au début de <l'année [1998]$_C$>.

+ <Le développement [o4a]$_X$> de <sa [o1]$_C$> présence en France <l' [o4a]$_I$> amènera <se [o1]$_C$> faire coter à Paris <le 12 juin [12 juin 1998]$_C$> prochain.

C'est <un nouveau groupe [o5]$_X$> Allianz <qui [o5]$_I$> naîtra d'<ici [12 juin 1998]$_I$> à <la fin de l'année [decembre 1998]$_{C,X}$>.

< [o5-rel-o1]$_{R-ID}$>

<L'assureur [o1]$_C$> allemand, <qui [o1]$_C$> consolide <ses [o1]$_C$> 51 % des AGF depuis <le 1er avril [1 avril 1998]$_C$>, considère que <cette prise [o2]$_C$> de contrôle <lui [o1]$_C$> confère « une très forte position dans <le secteur [o6]$_X$> de l'assurance mondiale, avec un pied particulièrement solide dans <notre marché [o8]$_X$> domestique qu'est <l'Europe [o8]$_\emptyset$> », comme l'explique < <son [o1]$_C$> président [o3a]$_X$>, Hennig Schulte-Noelle.

< [o3a-pde-o1]$_T$>
< [o1-mde-o6]$_{s30}$>
< [o8-pde-o6]$_{P-M}$>

Tout en <se [o3a]$_I$> félicitant [5] de < <son [o1]$_C$> acquisition [o2]$_C$>, <Al-

[5] L'annotateur explique son annotation en disant que « se féliciter » est typiquement une action humaine, et qu'il s'agit donc ici du président plutôt que de la société.

lianz [o1]$_C$> n'en rappelle pas moins <ses [o1]$_C$> objectifs.

Contre-pied d'<Axa [o9]$_X$>.

< [o9-mde-o6]$_{s32}$>

D'<ici [1998]$_I$> à 2000, <le résultat [o11]$_X$> net des AGF devra être porté à 5,5 milliards de francs, permettant à <sa [o1]$_I$> maison mère de constater un retour sur investissement de 9 %.

Pour 1998, <l'acquisition [o2]$_C$> des AGF doit permettre à <Allianz [o1]$_C$> de porter <son [o1]$_C$> chiffre d'affaires à 107 milliards de <deutsche marks [o3b]$_X$> (358 milliards de francs [o3b]$_C$) et de constater une croissance à deux chiffres de < <son [o1]$_C$> bénéfice [o12]$_X$> net.

< [o11-mde-o12]$_{s42}$> [6]

Mais pour <l'heure [o10]$_C$>, la priorité est à <la « digestion » [o13]$_X$> de <cette opération [o2]$_C$> d'<ici [o10]$_C$> à <fin 1999 [decembre 1999]$_T$> — ce <qui [o13]$_C$> passe par la détermination, <cette année [1998]$_C$>, des structures, des cadres dirigeants et des plans de développement dans chaque pays.

Au passage, <l'assureur [o1]$_C$> allemand rappelle <sa [o1]$_C$> volonté de décentralisation et, comme pour prendre le contre-pied de son grand concurrent Axa, affirme qu'<il [o3a]$_I$> ne voit pas l'intérêt de coiffer ses sociétés nationales d'une marque mondiale.

A <l'heure [o10]$_C$> actuelle, aucune décision n'a été prise quant à <d'éventuelles cessions [o14]$_X$> d'actifs, et notamment <les 25 % [o15]$_X$> de la Coface, comme le demande Bruxelles.

< [o15-mde-o14]$_C$>

Face à <cette nouvelle donne [o15]$_{s54}$>, <Allianz [o1]$_C$> réfléchit à la façon de parvenir à une présentation commune et aux économies possibles entre Euler et Hermès, sans pour autant créer une structure faîtière.

Enfin, toujours en France, <Hennig Schulte-Noelle [o3a]$_C$> a réaffirmé <son [o3a]$_C$> intention d'entrer jusqu'à 10 % dans le capital du Crédit Lyonnais.

D'un point de vue stratégique, <cette année [1998]$_C$> devrait marquer, selon <Allianz [o1]$_C$>, le démarrage de <sa [o1]$_C$> coopération dans la gestion d'actifs avec <la Dresdner Bank [o16]$_X$>, pour <laquelle [o16]$_\emptyset$> la répartition future des compétences reste en négociation.

En <assurance [o22(G)]$_X$>, <Allianz [o1]$_C$> discute avec la nouvelle Bayerische und Vereinsbank une répartition des partenariats entre <lui [o1]$_C$> et

[6] Le bénéfice prend en compte le résultat des AGF.

<Ergo [o17]$_X$>.

A ce sujet, et pour tenter de briser le soupçon de cartel porté sur le marché allemand, <Allianz [o1]$_C$> rappelle que <ses [o1]$_C$> 10 % d'<Ergo [o17]$_C$> ne sont pas stratégiques et que <les deux compagnies [o18]$_X$> restent « <des concurrentes [o18]$_\emptyset$> acharnées ».

< [{o1,o17}-mde-o18]$_C$>

Du reste, <Allianz [o1]$_C$> (qui n'a connu en 1997 qu'une hausse de ses primes de 0,4 % sur le marché automobile allemand en croissance moyenne de 1,3 %), affirme qu'<il [o1]$_C$> « continuera à prendre des mesures pour maintenir, sinon augmenter, <sa [o1]$_C$> part de marché sur ce segment très concurrentiel ».

Autant dire que la guerre des tarifs en Allemagne n'est pas terminée.

En <1997 [o21]$_X$>, <l'assureur [o1]$_C$>, qui a publié <un bénéfice [o19]$_X$> définitif de 2,7 milliards de marks (+ 20,5 %), a, pour la première fois, affiché <l'an [o21]$_C$> dernier <un bénéfice [o20]$_X$> technique positif de 182 millions.

< [o20-mde-o19]$_{s43}$>

Mais en <assurance-dommages [o23]$_X$>, le chiffre reste négatif de 1,3 milliard.

< [o23-mde-o22]$_{M-R}$>

<Mouvement [o26]$_X$> d'ouverture.

A l'heure où <il [o1]$_C$> veut compter parmi les cinq leaders mondiaux de l'assurance et où <il [o1]$_C$> prépare son entrée à <la Bourse de Paris [o25]$_X$> (le 12 juin prochain), <Allianz [o1]$_C$> a décidé de <s' [o1]$_C$>essayer à la transparence.

<Il [o1]$_C$> publiera donc des comptes aux <normes [o24]$_X$> internationales IASC à <la fin de l'année [decembre 1998]$_{C,C}$>, bien que <celles-ci [o24]$_C$> n'aient pas encore d' adaptation spécifique au <secteur [o25]$_I$>.

Surtout, <le groupe [o1]$_C$> a, « avec deux ans d'avance sur les obligations réglementaires », publié <ses [o1]$_C$> réserves cachées, l'équivalent de nos plus-values latentes, d'un montant de 87,7 milliards de <marks [o4b]$_X$> (près de 300 milliards de francs [o4b]$_C$).

Mais les observateurs sont impatients de voir <ce mouvement [o26]$_C$> d'ouverture <s' [o26]$_C$>étendre encore dans la comptabilité d'<Allianz [o1]$_C$>, qui se refusait hier à livrer une estimation de <sa [o1]$_C$> rentabilité (12,4 % en <1997 [o21]$_\emptyset$>, aux normes allemandes) fondée sur cette nouvelle valorisation de <ses [o1]$_C$> fonds propres.

A.6 Annotation réponse 5

Texte B

<[vendredi 29 mai 1998] >

<La BNP [o1]$_X$> <réorganise [o11]$_X$> < <son [o1]$_C$> pôle [o2]$_X$> de <financements [o5]$_X$> spécialisés...

Comme <elle [o1]$_C$> <l' [o6(P)]$_C$> avait laissé entendre en <février [fevrier 1998]$_C$>,< <la BNP [o1]$_C$> a décidé de <rapprocher [o11]$_\emptyset$> < <ses [o1]$_C$> filiales [o2]$_C$> de <crédits [o5]$_C$> spécialisés, <le Crédit Universel [o13]$_X$> (<biens [o9]$_X$> d'équipement pour les entreprises, les professionnels et les particuliers) et <BNP Bail [o12]$_X$> (<crédit-bail [o10]$_X$> mobilier et immobilier, location) [o6]$_X$>.

<[{o9,o10}-mde-o5]$_{C,C}$>
(<[{o12,o13}-mde-o2]$_T$>)

<Le nouvel ensemble [o2]$_C$>, baptisé BNP Lease, affiche, sur la base de <[o3]$_{s8}$> 1997, un produit net bancaire de 1,7 milliard de <francs [franc, o7]$_X$> et un résultat avant impôts de 700 millions <[franc, o7]$_C$>.

< <Sa [o2]$_C$> production [o3]$_X$> annuelle < <s' [o3]$_C$> élève [s'[]elever, o8]$_X$> à 20 milliards de <francs [o7]$_\emptyset$> et < <ses [o2]$_C$> encours [o4]$_X$> <[<s'[o4]$_\emptyset$> elever, o8]$_C$> à 54 milliards <[franc, o7]$_C$>.

<Le nouvel ensemble [o2]$_C$>, détenu à 100 % par <la banque [o1]$_C$>, sera présidé par <Claude Porcherot [o14]$_X$>, tandis que <Jean-René Brunon [o15]$_X$> <en [o2]$_C$> assurera la direction générale.

<[o2-mde-o1]$_T$>
<[o14-pde-o2]$_T$> ?
<[o15-pde-o2]$_T$> ?

Texte G

<[mardi 26 mai 1998,o0] >

<Guigou [o11]$_X$> visite <les locaux [o3]$_X$> du <pôle [o1]$_X$> financier

Le ban et l'arrière-ban de <la magistrature [o27] ?$_X$> parisienne étaient présents, <hier [25 mai 1998]$_C$>, aux côtés d'<Elisabeth Guigou [o11]$_C$> pour

visiter <les futurs locaux [o3]$_C$> du <pôle [o1]$_C$> financier parisien.

<[o1-mde-o27]$_{s15}$> ?

Après des mois de <discussion [o12]$_X$>, <le choix [o2]$_X$> définitif <s' [o2]$_C$> est porté, comme prévu, sur <le siège [o3]$_C$> historique du journal le Monde, <qui[o3] se[o3] trouve>$_\emptyset$ rue des Italiens (La Tribune du <19 mai [19 mai 1998]$_C$>), au coeur du quartier parisien de <la finance [o30]$_X$>.

<[o12-rel-o2]$_{s18}$>
<[o1-rel-o30]$_{s9}$>

<Cet immeuble [o3]$_C$> luxueux, complètement réaménagé, accueillera sur <6.400 mètres [o13]$_X$> carrés, d'<ici [o0]$_C$> à la fin de <l'année [1998]$_C$>, <274 <magistrats [o5]$_X$> et <fonctionnaires [o6]$_X$> [o26]$_X$>, plus <une trentaine [o7]$_X$> d'assistants spécialisés des <Finances [o30]$_{s27}$> et de <la Banque de France [o32]$_X$>.

<[{o5,o6}-mde-o26]$_T$>
<[o13-pde-o3]$_{s19}$>
<[o32-rel-o30]$_{s20}$>

<Ils [o8]$_X$> auront à <leur [o8]$_C$> disposition <quelque 23 mètres [o9]$_X$> carrés <par personne [o8(D)]$_C$>, contre pratiquement <moitié [o10]$_X$> moins auparavant au <Palais [o28]$_X$> de justice.

<[o9-rel-o10]$_C$>
<[o26,o7-mde-o8]$_C$>
<[o9-mde-o13]$_{M-ID}$> ?

<Montant [o15]$_X$> du loyer : <21,6 millions [o15]$_C$> de <francs [o24]$_X$> par an <auxquels [o15]$_C$> <s' [o14]$_C$> ajoutent <15 millions [o14]$_X$> <[o24]$_C$> de travaux spécifiques pour sécuriser <les lieux [o3]$_C$> <que [o3]$_I$> prend en charge le propriétaire de <l'immeuble [o3]$_C$>, la Bred.

Outre <des effectifs [o8 ?]$_C$> regroupés et supplémentaires, <le pôle [o1]$_C$> financier bénéficiera d'<équipements [o16]$_X$> informatiques.

<Les magistrats [o5]$_C$> auront notamment à <leur [o5]$_C$> disposition <des logiciels [o25]$_X$> d'instruction assistée par ordinateur .

<[o25-mde-o16]$_C$>

<Actuellement [o0]$_C$>, seuls <les juges [o33]$_X$> Eva Joly et Jean-Pierre Zanotto <en [o25]$_{ID-d}$ [o16]$_\emptyset$> disposent à <la galerie [o29]$_X$> financière de Paris .

<[o29-mde-o28]$_{s12}$> ?
<[o29-mde-o27]$_{s16}$> ?
<[o29-rel-o30]$_{s17}$>

<[o33-pde-o27]$_{P-M}$>

A <ceux [o17(I)]$_X$> <qui [o17]$_C$> <se [o17]$_C$> montrent réticents sur la mise en place de <ce pôle [o1]$_C$> financier, <le ministre [o11] ?$_C$> de la Justice a tenu à répéter que, « qu'<on [o18(I)]$_X$,[o17]$_\emptyset$,[o1]$_\emptyset$> <le [o19(P)]$_C$, [o20(P)]$_\emptyset$> veuille ou non, <on [o18]$_C$> n'échappera pas à <un besoin [o20]$_\emptyset$ de <spécialisation [o19]$_X$> croissante des <magistrats [o21(G)]$_X$> en matière d'information économique et financière ».

<[o5-mde-o21]$_C$>

<Cette annexe [o1]$_C$> parisienne du <Palais [o27]$_I$> de justice dédiée aux <dossiers [o31]$_X$> financiers devrait rapidement être suivie d'<autres pôles [o22]$_X$> <[o30] ?$_{s28}$> en province.

<[o11-pde-o27]$_{s21}$>
<[o1-dde-o22]$_C$>
<[o1-mde-o27]$_{s15}$>
<[o31-rel-o30]$_{s22}$>

<Le premier [o23]$_X$> sur <la liste [o22 ?]$_{ID-R}$> du gouvernement est le pôle corse.

<[o23-mde-o22]$_C$>

Texte A

<[vendredi 29 mai 1998, o0] >

Fort du <rachat [o7]$_X$> des <AGF [o2]$_X$>, <Allianz [o1]$_X$> présente < <son [o1]$_C$> nouveau visage [o19]$_X$>

<[o2-mde-o1]$_\emptyset$>

+ <L'assureur [o1]$_C$> allemand a engagé <un tournant [o19 ?]$_{s59}$> stratégique avec <le rachat [o7]$_C$> des <AGF [o2]$_C$> au début de <l'année [1998, o3]$_C$>.

<[o19-rel-o7]$_T$> ?

+ <Le développement [o7]$_{ID-R}$> de <sa [o1]$_C$> présence en <France [o9]$_X$> <l' [o1]$_C$> amènera <se [o1]$_C$> faire coter à <Paris [o17]$_X$> <le 12 juin [12 juin 1998]$_C$> prochain.

<[o17-rel-o9]$_\emptyset$>

C'est <un nouveau groupe [o12]$_X$> Allianz <qui [o12]$_I$> naîtra d'<ici [o0]$_C$>

à la fin de <l'année [o3]$_C$>.

<[o2-mde-o12]$_{s44}$>
<[o1-rel-o12]$_{R-ID}$>
<[o19-rel-o12]$_{s45}$>

<L'assureur [o1]$_C$> allemand, <qui [o1]$_C$> consolide <ses [o1]$_C$> 51 % des <AGF [o2]$_C$> depuis <le 1er avril [01 avril 1998]$_C$>, considère que < <cette prise [o7]$_C$> de contrôle <lui [o1]$_C$> confère « une très forte position dans le secteur de <l'assurance [o18]$_X$> mondiale, avec un pied particulièrement solide dans < <notre [o1]$_C$> marché [o5]$_X$> domestique <qu' [o5]$_C$> est l'Europe » [o6]$_X$>, comme <l' [o6(P)]$_C$> explique < <son [o1]$_C$> président [o13]$_X$>, Hennig Schulte-Noelle.

(<[o13-pde-o1]$_T$>)

Tout en <se [o1]$_C$> félicitant de < <son [o1]$_C$> acquisition [o7]$_C$>, <Allianz [o1]$_C$> n'<en [o20 ?]$_{s51}$> rappelle pas moins < <ses [o1]$_C$> objectifs [o20]$_X$>.

Contre-pied d'<Axa [o8]$_X$>.

D'<ici [o0]$_C$> à 2000, <le résultat [o22]$_X$> net des <AGF [o2]$_C$> devra être porté à 5,5 milliards de <francs [o49]$_X$>, permettant à < <sa [o2]$_C$> maison [o1]$_C$> mère de constater <un retour [o23]$_X$> sur <investissement [o56]$_X$> de <9 % [o48]$_X$>.

<[o22-pde-o20]$_{P-M}$> ?
<[o23-pde-o20]$_\emptyset$> ?
<[o23-rel-o22]$_{s36}$>
<[o48-rel-o56]$_T$>
<[o7-rel-o56]$_{s46}$>

Pour <1998 [o3]$_T$>, <l'acquisition [o7]$_C$> des <AGF [o2]$_C$> doit permettre à <Allianz [o1]$_C$> de porter < <son [o1]$_C$> chiffre [o26]$_X$> d'affaires à <107 milliards [o25]$_X$> de <deutsche marks [o40]$_X$> (<358 milliards [o25]$_C$> de <francs [o49]$_\emptyset$>) et de constater <une croissance [o24]$_X$> à deux chiffres de <son [o1]$_C$> bénéfice net.

<[o24-pde-o20]$_{P-M}$> ?
<[o26-pde-o20]$_{P-M}$> ?
<[o26-rel-o24]$_{s47}$>

Mais pour <l'heure [o0]$_C$>, la priorité est à <la « digestion » [o27]$_X$> de <cette opération [o7]$_C$> d'<ici [o0]$_C$> à fin 1999 — ce <qui [o27]$_C$> passe par <la détermination [o28]$_X$>, <cette année [o3]$_C$>, des structures,<[o28]$_\emptyset$> des cadres dirigeants et <[o28]$_\emptyset$> des plans de développement dans <chaque pays

[o30(D),(I)]$_X$>.

Au passage, <l'assureur [o1]$_C$> allemand rappelle <sa [o1]$_C$> volonté de <décentralisation [o29]$_X$> et, comme pour prendre le contre-pied de < <son [o1]$_C$> grand concurrent [o8]$_C$> Axa, affirme qu'<il [o1]$_C$> ne voit pas l'intérêt de coiffer < <ses [o1]$_C$> sociétés [o31]$_X$> <nationales [o30(I)]$_{s61}$> d'une marque mondiale.

<[o29-pde-o20]$_{P-M}$>

A <l'heure [o0]$_C$> actuelle, <aucune décision [o46(D)]$_X$> n'a été prise quant à <d'éventuelles cessions [o10(G)]$_X$> d'actifs, et notamment <[o10]$_\emptyset$> <les 25 % [o11]$_X$> de <la Coface [o32]$_X$>, comme <le [o10(P)]$_C$, [o46(P)]$_\emptyset$> demande Bruxelles.

<[o11-rel-o10]$_{R-M}$>
<[o32-pde-o1]$_\emptyset$> ?

Face à <cette nouvelle donne [o10 ?]$_{s54}$>, <Allianz [o1]$_C$> réfléchit à la façon de parvenir à une présentation commune et aux économies possibles entre Euler et Hermès, sans pour autant créer une structure faîtière.

Enfin, toujours en <France [o9]$_C$>, <Hennig Schulte-Noelle [o13]$_C$> a réaffirmé < <son [o13]$_C$> intention [o41]$_X$> d'entrer jusqu'à <10 % [o35]$_X$> dans <le capital [o47]$_X$> du Crédit Lyonnais.

<[o41-pde-o20]$_{P-M}$>
<[o35-pde-o47]$_T$> ?

D'un point de vue stratégique, <cette année [o3]$_C$> devrait marquer, selon <Allianz [o1]$_C$>, le démarrage de < <sa [o1]$_C$> coopération [o33]$_X$> dans la gestion d'actifs avec <la Dresdner Bank [o14]$_X$>, pour <laquelle [o14]$_\emptyset$> la répartition future des compétences reste en négociation.

<[o33-pde-o20]$_{P-M}$>

En <assurance [o42]$_{s52}$>, <Allianz [o1]$_C$> discute avec <la nouvelle Bayerische und Vereinsbank [o16]$_X$> <une répartition [o15]$_X$> des partenariats entre <lui [o16 ?]$_I$> et <Ergo [o34]$_X$>.

<[o15-pde-o20]$_{P-M}$> ?

A <ce sujet [o15]$_C$>, et pour tenter de briser le soupçon de cartel porté sur le marché allemand, <Allianz [o1]$_C$> rappelle que <ses [o1]$_C$> 10 % d'<Ergo [o34]$_C$> ne sont pas stratégiques et que <les deux compagnies [o37]$_X$> restent « des concurrentes acharnées ».

<[{o1,o34}-mde-o37]$_C$>

Du reste, <Allianz [o1]$_C$> (<qui [o1]$_C$> n'a connu en 1997 qu'une hausse de <ses [o1]$_C$> primes de 0,4 % sur <le marché [o36]$_X$> automobile allemand

en croissance moyenne de 1,3 %), affirme qu'<il [o1]$_C$> « continuera à prendre <des mesures [o38]$_X$> pour maintenir, sinon augmenter, <sa [o1]$_C$> part de marché sur <ce segment [o36]$_C$> très concurrentiel ».

<[o38-pde-o20]$_{P-M}$>

Autant dire que la guerre des tarifs en Allemagne n'est pas terminée.

En 1997, <l'assureur [o1]$_C$>, <qui [o1]$_C$> a publié un bénéfice définitif de 2,7 milliards de <marks [o40]$_\emptyset$> (+ 20,5 %), a, pour la première fois, affiché <l'an [1997]$_C$> dernier un bénéfice technique positif de 182 millions <[o40]$_C$>.

Mais en <assurance-dommages [o39]$_X$>, le chiffre reste négatif de 1,3 milliard <[o40]$_C$>.

<[o39-mde-o42]$_{M-R}$>

<Mouvement [o51]$_X$> d'ouverture .

A <l'heure [o0]$_C$> <où [o0]$_C$> <il [o1]$_C$> veut compter parmi les cinq leaders mondiaux de <l'assurance [o42]$_X$> et <où [o0]$_C$> <il [o1]$_C$> prépare <son [o1]$_C$> entrée à <la Bourse de Paris [o17]$_C$> (<le 12 juin [o43]$_\emptyset$> prochain), <Allianz [o1]$_C$> a décidé de <s' [o1]$_C$> essayer à la transparence.

<Il [o1]$_C$> publiera donc <des comptes [o55]$_X$> aux <normes [o45]$_X$> internationales IASC à la fin de <l'année [o3]$_C$>, bien que <celles-ci [o45]$_C$> n'aient pas encore d' adaptation spécifique au <secteur [o42]$_C$>.

Surtout, <le groupe [o1]$_C$> a, « avec deux ans d'avance sur <les obligations [o54]$_X$> réglementaires », publié < <ses [o1]$_C$> réserves [o50]$_X$> cachées, l'équivalent de < <nos [o9 ?]$_C$> plus-values [o50]$_{s60}$> latentes, d'un montant de <87,7 milliards [o44]$_X$> de <marks [o40]$_\emptyset$> (<près de 300 milliards [o44]$_C$> de <francs [o49]$_\emptyset$>).

<[o54-rel-o45]$_{s48}$>

Mais les observateurs sont impatients de voir <ce mouvement [o51]$_C$> d'ouverture <s' [o51]$_C$> étendre encore dans <la comptabilité [o52]$_X$> d'<Allianz [o1]$_C$>, <qui [o52]$_I$> <se [o52]$_I$,[o1]$_\emptyset$> refusait <hier [1998-05-28]$_C$> à livrer une estimation de < <sa [o52]$_I$> rentabilité [o53]$_X$> (12,4 % en 1997, aux normes allemandes) fondée sur <cette nouvelle valorisation [?]$_\emptyset$> de < <ses [o52]$_I$,[o1]$_\emptyset$> fonds [o51]$_{s58}$> propres.

<[o51-pde-o1]$_T$>

A.7 Inventaire des liens observés

La présente section, la dernière de l'annexe contenant les données du test d'opérationalité présenté au chapitre 4, contient sous forme de tableaux un inventaire de tous les liens observés, soit par l'expert, soit par les annotateurs.

Chaque tableau rassemble les liens observés par rapport à un type de lien donné : identités de description, reprises de type paraphrase, expressions temporelles, etc. Pour la coréférence, les données sont en outre classées par types d'expressions. Enfin, pour la coréférence et les relations référentielles, des tableaux différents sont consacrés au relevé des observations qui correspondent à une observation de l'expert, d'une part, et au relevé des observations superflues, d'autre part.

Chaque ligne d'un tableau identifie un lien. La première colonne contient un identifiant qui permet des références croisées entre les tableaux et les annotations présentées plus haut : un simple nombre renvoie à un indice figurant dans l'annotation clé, un nombre précédé de « s » à un indice figurant dans une ou plusieurs des annotations réponses. La seconde colonne reproduit sous forme abrégée les expressions entre lesquelles le lien a été observé.

La colonne *ex* donne le type de l'expression-reprise considérée :

- ph = phrases ou propositions,
- np = noms propres,
- sn = syntagmes nominaux descriptifs,
- pr = pronoms autres que réfléchis et relatifs,
- po = déterminants possessifs,
- rl = pronoms relatifs,
- rf = pronoms réfléchis,
- e = ellipse,
- ad = adverbe.

La colonne *clé* donne le type du lien tel qu'observé par l'expert :

- d = identité de description,
- p = reprise de type paraphrase,
- t = expressions temporelle,
- ID = identité de dénotation (ou « coréférence »),
- R = relation `rel`,
- P = relation `partie-de`,
- M = relation `membre-de`,
- D = relation `distingué-de`.

Une case vide dans cette colonne correspond à une absence d'annotation dans la clé, donc à un ou plusieurs liens superflus dans les annotations réponses.

Les colonnes *1* à *5* donnent les prédicats d'évaluation pour chacune des anno-

tations réponses. Comme indiqué dans la section 4.2.2 du chapitre 4 (page 141), les prédicats d'évaluation sont ici plus détaillés que ceux qui sont utilisés pour le calcul des mesures de rappel et précision, cela pour permettre le calcul de la variance. La sémantique des valeurs trouvées dans ces colonnes est la suivante :

- C note une reponse jugée correcte au regard de la clé,
- une case vide note une absence complète de réponse,
- les symboles d, ID, ID', ID", R, P, M et D notent qu'un lien du type correspondant a été annoté par l'annotateur ; soit ce lien correspond à un lien d'un type différent dans la clé, soit c'est un lien superflu,
- les symboles ID, ID' et ID" distinguent en outre des annotations de lien de coréférence qui, pour une expression, n'expriment pas la même interprétation,
- les symboles I et I' notent des liens jugés incorrects, sauf pour les liens avec relation référentielle, pour lesquels les symboles précédents jouent ce rôle,
- les symboles I et I' distinguent deux annotations incorrectes qui n'expriment pas la même interprétation.

Enfin, la dernière colonne donne les prédicats d'évaluation pour l'annotation de l'« observateur idéal » issue de l'opinion majoritaire.

	Liens	*ex*	*clé*	*1*	*2*	*3*	*4*	*5*	*m*
13	700 millions → 1,7 milliards de F	sn	d	C				C	
17	∅ → s'élève	e	d	ID	C			C	C
18	54 milliards → 20 milliards de F	sn	d	C				C	
43	15 millions → 21,6 millions de F	sn	d	C			C	C	C
54	en → logiciels	pr	d	ID	ID	ID	ID	ID	ID
165	182 millions → 2,7 m. de marks	sn	d	C				C	
168	1,3 milliards → 2,7 m. de marks	sn	d	C				C	
184	celles-ci → les normes int. IASC	pr	ID	d	C	C	C	C	C

TAB. A.1 – Identité de description

	Liens	*ex*	*clé*	*1*	*2*	*3*	*4*	*5*	*m*
3	l'[avait laissé entendre] → racheter	pr	p	ID	C	ID	I	C	ID
59	le [veuille] → échappera	pr	p	ID	ID	ID	ID	C	ID
92	l'[explique] → confère forte position	pr	p	ID	ID	ID		C	ID
130	le [demande] → cession 25% de la C.	pr	p	ID	ID	ID		C	ID

TAB. A.2 – Reprises de type paraphrase

	Liens	*ex*	*clé*	*1*	*2*	*3*	*4*	*5*	*m*
4	en février → `[février 1998]`	sn	t	C	C	C	C	C	C
22	hier → `[25 mai 1998]`	ad	t	C	C	C	C	C	C
28	le 19 mai → `[19 mai 1998]`	sn	t	C	C		C	C	C
30	ici → `[26 mai 1998]`	ad	t	C	I	C	I'	C	C
31	l'année → `[1998]`	sn	t	I	C	C	C	C	C
52	Actuellement → `[26 mai 1998]`	ad	t			I		C	
72	l'année → `[1998]`	sn	t	C	C	C	C	C	C
76	le 12 juin prochain → `[12 juin 1998]`	sn	t	C	C	C	C	C	C
80	ici → `[29 mai 1998]`	ad	t	C	I		I	C	I
81	l'année → `[1998]`	sn	t	I	C	C	C	C	C
88	le 1er avril → `[1er avril 1998]`	sn	t	C	C	C	C	C	C
100	ici → `[29 mai 1998]`	ad	t	C	I	C	I	C	C
113	pour l'heure → `[29 mai 1998]`	sn	t	C	I	C	C	C	C
116	ici → `[29 mai 1998]`	ad	t	C	I	C	C	C	C
119	cette année → `[1998]`	sn	t	C	C	C	C	C	C
128	à l'heure actuelle → `[29 mai 1998]`	sn	t	C	I	C	C	C	C
139	cette année → `[1998]`	sn	t	C	C	C	C	C	C
164	l'an dernier → `[1997]`	sn	t	C	C	C	C	C	C
169	l'heure → `[29 mai 1998]`	sn	t	C	I			C	C
178	le 12 juin prochain → `[12 juin 1998]`	sn	t	C	C	C			C
182	l'année → `[1998]`	sn	t	I	C	C	C	C	C
183	la fin de l'année → la fin de l'année	sn	t	I	C	C	C		C
197	hier → `[28 mai 1998]`	ad	t		C			C	

TAB. A.3 – Expressions temporelles

	Liens	*ex*	*clé*	*1*	*2*	*3*	*4*	*5*	*m*
3	l'[avait laissé entendre] → racheter	pr	p	ID	C	ID	I	C	ID
17	∅ → s'élève	e	d	ID	C			C	C
s8	`<ellipse>` [1997] → sa production	sn						ID	
39	le loyer → les locaux	sn	R				ID		
51	logiciels → équipements	sn	M	P	ID	C	P	C	P
54	en → logiciels	pr	d	ID	ID	ID	ID	ID	ID
59	le [veuille] → échappera	pr	p	ID	ID	ID	ID	C	ID
61	les magistrats (SP) → les mag. (G)	sn	M	ID	P		C	C	C
66	la liste → autres pôles	sn	R				ID	ID	
s24	specialisation → le pôle	sn				ID			
s25	mag. + fonc. + assist. → le pôle	sn					ID		
s26	les côtés d' EG → la magistrature	sn					ID		
s27	les Finances → la finance	sn		ID				ID	
s28	`<ellipse>` → la finance	sn						ID	
77	le dév. de sa présence → le rachat	sn	R					ID	
92	l'[explique] → confère forte position	pr	p	ID	ID	ID		C	ID
130	le [demande] → cession 25% de Cof.	pr	p	ID	ID	ID		C	ID
167	le chiffre → bénéf. tech. de 182 m.	sn	R		ID				
s41	ses 51 % des AGF → Allianz	sn				ID			
s50	aucune décision → évent. cessions	sn			ID				
s51	[n'] en [rappelle] → `[divers]`	pr			ID'	ID"		ID	
s52	[en] assurance → l'assurance	sn			ID			ID	
s53	son bénéfice net → rés. net des AGF	sn			ID				
s54	nouvelle donne → `[divers]`	sn			ID'	ID"	ID	ID	ID
s55	notre marché domest. → les AGF	sn				ID			
s56	ses primes → sa part de marché	sn				ID			
s57	ses objectifs → le rachat des AGF	sn				ID			
s58	ses fonds propres → ce mouv. d'ouv.	sn						ID	
s59	tournant strat. → son nouv. visage	sn						ID	
s60	nos plus-values lat.→ ses rés. cach.	sn						ID	
s61	[sociétés] nationales → chaque pays	sn						ID	

TAB. A.4 – Coréférence. Ensemble des annotations superflues.

	Liens	*ex*	*clé*	*1*	*2*	*3*	*4*	*5*	*m*
86	qui consolide ses 51% → le rachat	ph	ID						
177	entrée à la B. de P. → se faire coter à P.	ph	ID						
191	publie réserves → Mouv. d'ouv.	ph	ID			C			

TAB. A.5 – Coréférence. Phrases ou propositions.

	Liens	*ex*	*clé*	*1*	*2*	*3*	*4*	*5*	*m*
7	ses filiales → son pôle	sn	ID	C	C		C	C	C
8	crédits spéc. → financement spéc.	sn	ID		C			C	
12	le nouvel ensemble → ses filiales	sn	ID	C	C	M	C	C	C
19	le nouvel ensemble → le nouvel ens.	sn	ID	C	C	C	C	C	C
20	la banque → la BNP	sn	ID	C	C	C	C	C	C
24	les futurs locaux du p. → les loc. du p.	sn	ID	C	C	C	C	C	C
25	le pôle fin. parisien → le pôle fin.	sn	ID	C	C		C	C	C
27	le siège du Monde → les futurs locaux	sn	ID	C	C	C	C	C	C
29	Cet immeuble → le siège du Monde	sn	ID	C	C	C	C	C	C
35	quelque 23 m^2 par pers. → 6400 m^2	sn	ID	R	P	M		M	M
36	par personne → Ils	sn	ID	C	C	M	I	C	C
40	21,6 mill. de f. → Montant du loyer	sn	ID					C	
44	les lieux → Cet immeuble	sn	ID	C	C	C	C	C	C
46	l'immeuble → les lieux	sn	ID		C	C	C	C	C
47	des effectifs → 274 m. et f. + assist.	sn	ID		C		I	C	C
48	le pôle financier → le pôle fin. parisien	sn	ID	C	C	C	C	C	C
49	Les magistrats → [274] magist. [et f.]	sn	ID	C	C	C	C	C	C
57	ce pôle financier → le pôle financier	sn	ID	C	C	C	C	C	C
58	le ministre de la Justice → E. Guigou	sn	ID	C	C		C	C	C
62	Cette annexe → ce pôle financier	sn	ID	I	I	I	I	C	I
63	le Palais de justice → le Palais de just.	sn	ID		I			I'	
69	L'assureur allemand → Allianz	sn	ID	C	C		C	C	C
70	le rachat des AGF → le rachat	sn	ID	C	C		C	C	C
82	L'assureur allemand → Allianz	sn	ID	C	C	I	C	C	C
87	cette prise de contrôle → le rachat	sn	ID	C	C	I	C	C	C
97	son acquisition → le rachat des AGF	sn	ID	C	I	I'	C	C	C
103	sa maison mère → Allianz	sn	ID		C			C	
105	l'acquisition des AGF → le rachat	sn	ID	C	C	C	C	C	C
109	358 m. de francs → 107 m. de DM	sn	ID	R			C	C	C
115	cette opération → le rachat des AGF	sn	ID	C	C	C	C	C	C
120	l'assureur allemand → Allianz	sn	ID	C	C	I	C	C	C
123	le c.-pied de gd conc. → C-p. d'Axa	sn	ID		C				
146	ce sujet → une répartition des part.	sn	ID	C	C	C		C	C
151	les deux compagnies → Allianz, Ergo	sn	ID	C	C	I	C	C	C
159	ce segment → le marché auto. all.	sn	ID		C	C		C	C
162	l'assureur → Allianz	sn	ID	C	C		C	C	C
172	l'assurance → le secteur de l'ass. mond.	sn	ID	C					
176	la Bourse de Paris → Paris	sn	ID					C	
185	le secteur → le secteur de l'ass. mondiale	sn	ID		C	C	I	C	C
186	le groupe → Allianz	sn	ID	C	C	C	C	C	C
189	près de 300 m. de F → 87,7 m. de DM	sn	ID	R			C	C	C
192	ce mouv. d'ouverture → Mouv. d'ouv.	sn	ID	C	C		C	C	C

TAB. A.6 – Coréférence. Syntagmes nominaux descriptifs

	Liens	*ex*	*clé*	*1*	*2*	*3*	*4*	*5*	*m*
5	la BNP → La BNP	np	ID	C	C		C	C	C
23	Elisabeth Guigou → Guigou	np	ID	C	C			C	C
71	les AGF → les AGF	np	ID	C	C			C	C
78	un nouveau groupe Allianz → Allianz	np	ID				R	R	
85	les AGF → les AGF	np	ID	C	C			C	C
98	Allianz → Allianz	np	ID	C	C	C	C	C	C
101	les AGF → les AGF	np	ID	C	C			C	C
106	les AGF → les AGF	np	ID	C	C			C	C
107	Allianz → Allianz	np	ID	C	C	C	C	C	C
124	son gd concurrent Axa → Axa	np	ID	C	C			C	C
131	Allianz → Allianz	np	ID	C	C		C	C	C
135	France → France	np	ID					C	
136	Hennig Schulte-Noelle → son président	np	ID		C		C	C	C
140	Allianz → Allianz	np	ID	C	C		C	C	C
143	Allianz → Allianz	np	ID	C	C		C	C	C
148	Allianz → Allianz	np	ID	C	C		C	C	C
150	Ergo → Ergo	np	ID	C	C		C	C	C
153	Allianz → Allianz	np	ID	C	C		C	C	C
179	Allianz → Allianz	np	ID	C	C		C	C	C
194	Allianz → Allianz	np	ID	C	C		C	C	C

TAB. A.7 – Coréférence. Noms propres.

	Liens	*ex*	*clé*	*1*	*2*	*3*	*4*	*5*	*m*
15	s'[élève] → Sa production	rf	ID	C	C		C	C	C
26	s'[est porté] → le choix	rf	ID	C	C		C	C	C
42	s'[ajoutent] → 15 millions de travaux	rf	ID	I	C		I	C	I
56	se [montrent réticents] → ceux qui	rf	ID	C	C		C	C	C
75	se [faire coter] → l'assureur allemand	rf	ID	C	C	I	C	C	C
95	se [félicitant] → Allianz	rf	ID	C	C	C	I	C	C
180	s'[essayer] → Allianz	rf	ID	C	C	C	C	C	C
193	s'[étendre] → ce mouvement d'ouverture	rf	ID	C	C	C	C	C	C
196	se [refusait] → Allianz	rf	ID	C	C	C		I	C

TAB. A.8 – Coréférence. Pronoms réfléchis.

	Liens	*ex*	*clé*	*1*	*2*	*3*	*4*	*5*	*m*
41	auxquels [s'ajoutent] → 21,6 millions	rl	ID	C	C	C	C	C	C
45	que [prend en charge] → 15 millions	rl	ID	I	C	C		I	I
55	qui [se montrent réticent] → ceux	rl	ID	C				C	
79	qui [naîtra] → un nv groupe Allianz	rl	ID	I	I	C	I	I	I
83	qui [consolide] → l'assureur allemand	rl	ID	C	C	I	C	C	C
91	qu'[est l'Europe] → notre marché	rl	ID	C	C	I		C	C
118	qui [passe] → ce [la « digestion »]	rl	ID	C			C	C	C
154	qui [n'a connu] → Allianz	rl	ID	C	C	C		C	C
163	qui [a publié] → l'assureur	rl	ID	C	C	I		C	C
170	où [il veut] → l'heure	rl	ID			I		C	
173	où [il prépare] → l'heure	rl	ID		I	I'		C	
195	qui [se refusait] → Allianz	rl	ID	C		C		I	C

TAB. A.9 – Coréférence. Pronoms relatifs.

	Liens	*ex*	*clé*	*1*	*2*	*3*	*4*	*5*	*m*
1	son [pôle] → la BNP	po	ID	C	C		C	C	C
6	ses [filiales] → la BNP	po	ID	C	C		C	C	C
14	Sa [production] → le nv. ensemble	po	ID	C	I	C	C	C	C
16	Ses [encours] → le nouvel ensemble	po	ID	C	I	C	C	C	C
34	leur [disposition] → Ils	po	ID	C	C	C	I	C	C
50	leur [disposition] → Les magistrats	po	ID	C	C	C	C	C	C
68	son [nouveau visage] → Allianz	po	ID	C		I	C	C	C
73	sa [présence] → l'assureur allemand	po	ID	C	C	I	C	C	C
84	ses [51% des AGF] → l'assureur all.	po	ID	C	C	I	C	C	C
90	notre [marché] → l'assureur allemand	po	ID	C	C			C	C
93	son [président] → l'assureur allemand	po	ID	C	C	I	C	C	C
96	son [acquisition] → Allianz	po	ID	C	C	C	C	C	C
99	ses [objectifs] → Allianz	po	ID	C	C	C	C	C	C
102	sa [maison-mère] → les AGF	po	ID	C	C	C	I	C	C
108	son [chiffre d'affaires] → Allianz	po	ID	C	C	C	C	C	C
112	son [bénéfice net] → Allianz	po	ID	C	C	C	C	C	C
121	sa [volonté] → l'assureur allemand	po	ID	C	C	I	C	C	C
125	son [grand concurrent] → l'assureur all.	po	ID	C	C	I		C	C
127	ses [sociétés] → l'assureur allemand	po	ID	C	C	I		C	C
137	son [intention] → H. Schulte-Noelle	po	ID	C	C	C	C	C	C
142	sa [coopération] → Allianz	po	ID	C	C	C	C	C	C
149	ses [10% d'Ergo] → Allianz	po	ID	C	C	C	C	C	C
155	ses [primes] → Allianz	po	ID	C	C	C		C	C
158	sa [part de marché] → Allianz	po	ID	C	C	C	C	C	C
175	son [entrée à la Bourse] → Allianz	po	ID	C	C	C		C	C
187	ses [réserves cachées] → le groupe	po	ID	C	C		C	C	C
188	nos [plus-values latentes] → `[français]`	po	ID		C			C	
198	sa [rentabilité] → Allianz	po	ID	C	C	C	C	I	C
199	ses [fonds propres] → Allianz	po	ID	C	C		C	I	C

TAB. A.10 – Coréférence. Déterminants possessifs.

	Liens	*ex*	*clé*	*1*	*2*	*3*	*4*	*5*	*m*
2	elle → la BNP	pr	ID	C	C	C	C	C	C
21	en → Le nouvel ensemble	pr	ID	C	C	C		C	C
33	Ils → 274 m. et f. + assistants	pr	ID	C	C	C	I	C	C
60	on → on	pr	ID		C		C	C	C
74	l'[amènera] → l'assureur allemand	pr	ID	C	C	I	I'	C	C
89	lui [confère] → l'assureur allemand	pr	ID	C	C	I	C	C	C
117	ce [qui passe] → la digestion	pr	ID	C	C	C			C
126	il [ne voit pas] → l'assureur all.	pr	ID	C	C	I	I'	C	C
145	lui → Allianz	pr	ID	C	C	C	C	I	C
157	il [continuera] → Allianz	pr	ID	C	C	C	C	C	C
171	il [veut compter] → Allianz	pr	ID	C	C	C	C	C	C
174	il [prépare] → Allianz	pr	ID	C	C	C	C	C	C
181	Il [publiera] → Allianz	pr	ID	C	C	C	C	C	C
184	celles-ci → les normes int. IASC	pr	ID	d	C	C	C	C	C
3	l'[avait laissé entendre] → racheter	pr	p	ID	C	ID	I	C	ID
54	en → logiciels	pr	d	ID	ID	ID	ID	ID	ID
59	le [veuille] → échappera	pr	p	ID	ID	ID	ID	C	ID
92	l'[explique] → confère forte position	pr	p	ID	ID	ID		C	ID
130	le [demande] → cession 25% de la C.	pr	p	ID	ID	ID		C	ID
s51	[n'] en [rappelle] → `[divers]`	pr			ID'	ID"		ID	

TAB. A.11 – Coréférence. Autres pronoms.

	Liens	*ex*	*clé*	*1*	*2*	*3*	*4*	*5*	*m*
9	biens d'équipement → crédits spéc.	sn	M		R			C	
10	crédit-bail → crédits spécialisés	sn	M		R			C	
11	location → crédits spécialisés	sn	M		R				
32	[274] magistrats → magistrature	sn	M	C	C				
37	moitié moins → 23 m^2 par pers.	sn	R	P	C		C	C	C
38	auparavant → d'ici à fin d année	ad	R						
39	le loyer → les locaux	sn	R				ID		
51	logiciels → équipements	sn	M	P	ID	C	P	C	P
53	les juges EJ et JPZ → la magistrature	np	M		P			P	
61	les magistrats (SP) → les mag. (G)	sn	M	ID	P		C	C	C
64	autres pôles → le pôle financier	sn	D	C	C	M	C	C	C
65	Le premier → autres pôles	sn	M	C	C	C	C	C	C
66	la liste → autres pôles	sn	R				ID	ID	
67	E. Guigou → le gouvernement	sn	M		C		C		
77	le dév. de sa présence → le rachat	sn	R					ID	
94	notre marché dom. → le sect. de l'ass.	sn	M				P		
104	résultat AGF à 5,5m. → ses objectifs	ph	M					P	
110	chiffre d'aff. à 107m. → ses objectifs	ph	M					P	
111	croissance bénéf. net → ses objectifs	ph	M					P	
114	la digestion de opér. → ses objectifs	sn	M						
122	sa volonté de décentralis. → objectifs	sn	M		C			P	
129	les 25% de Coface → actifs	sn	M			C	C	R	C
132	prés. commune et économ. → objectifs	ph	M						
133	Euler → ses sociétés nationales	np	M						
134	Hermès → ses sociétés nationales	np	M						
138	entrer ds capital Crédit L. → objectifs	ph	M		C			P	
141	démarrage coop. avec DB → objectifs	sn	M					P	
144	répartition des parten. → objectifs	sn	M					P	
147	cartel → les deux compagnies	sn	R						
152	le marché allemand → marché dom.	sn	M						
156	le marché auto. all. → le marché all.	sn	M			C			
160	augmenter part de marché → objectifs	ph	M					P	
161	des mesures → la guerre des tarifs	sn	R						
166	assurance-dommages → assurance	sn	R				M	M	
167	le chiffre → un bénéf. tech. de 182 m.	sn	R		ID				
190	réserves cachées → comptes	sn	M						
200	la comptabilité → comptes	sn	R						
201	nouv. valorisation → publie réserves	ph	R						
202	ses rés. cachées → ses fonds propres	sn	P						

TAB. A.12 – Relations référentielles possibles

	Liens	*ex*	*clé*	*1*	*2*	*3*	*4*	*5*	*m*
12	le nouvel ensemble → ses filiales	sn	ID	C	C	M	C	C	C
s1	la BNP → le nouvel ensemble	sn				M			
s2	un résultat → un produit net	sn			P				
s3	un produit net → sa production	sn			P				
s4	ses encours → sa production	sn			R				
s5	un produit net → (sur) la base de 1997	sn					R		
s6	un résultat → (sur) la base de 1997	sn					R		
s7	la base → le nouvel ensemble	sn					R		
35	quelque 23 m^2 par pers. → 6400 m^2	sn	ID	R	P	M		M	M
36	par personne → Ils	sn	ID	C	C	M	I	C	C
s9	la finance → le pôle financier	sn		R				R	
s10	ordinateur → équipements	sn		P					
s11	logiciel/équipements → ordinateur	sn			M				
s12	la galerie → le Palais de justice	sn			P			M	
s13	ceux qui → Ils	pr			P				
s14	Ils → on	pr			P				
s15	le pôle → la magistrature	sn						M	
s16	la galerie fin. → la magistrature	sn						M	
s17	la galerie fin. → les Finances	sn						R	
s18	discussions → le choix	sn						R	
s19	6400 m^2 → les locaux	sn						P	
s20	Banque de France → Finances	sn						R	
s21	Guigou → Palais de j. | magistrature	sn						P	
s22	les dossiers financiers → la finance	sn						R	
s23	(après) des mois → hier	sn		R					
78	un nouveau groupe Allianz → Allianz	np	ID				R	R	
109	358 m. de francs → 107 m. de DM	sn	ID	R			C	C	C
189	près de 300 m. de F → 87,7 m. de DM	sn	ID	R			C	C	C
s29	hier → 29 mai 1998 [t de l'article]	ad		D					
s30	Allianz → assurance mondiale	sn		M			M		
s31	AGF → assurance mondiale	sn		M					
s32	Axa → assurance mondiale	sn		M			M		
s33	nouv. valorisation → mouv. ouverture	sn		P					
s34	contre-pied d'Axa → objectifs	sn			M				
s35	mouv. d'ouverture → objectifs	sn			M				
s36	retour investiss. → résultat net	sn			P			R	
s37	résultat net → chiffre d'affaire	sn			P				
s38	rentabilité → chiffre d'affaire	sn			R				
s39	9% → 5,5 milliards de francs	sn				R			
s40	rentabilité → comptabilité	sn				R			
s42	résultat net AGF → bénéf. net Allianz	sn					M		
s43	bénéf. technique → bénéf. définitif	sn			D		M		
s44	AGF → un nouveau groupe	sn						M	
s45	tournant stratégique → nouveau groupe	sn						R	
s46	rachat → investissement	sn						R	
s47	chiffre d'affaire → croissance	sn						R	
s48	obligations → normes int. IASC	sn						R	
s49	du reste → [à] ce sujet	sn				D			

TAB. A.13 – Relations référentielles superflues

Annexe B

Exemple d'analyse syntaxique

La présente annexe donne l'analyse syntaxique obtenue avec le système XIP pour le texte suivant :

> Le groupe Paribas va céder la participation de 25 % qu'il détient dans la banque d'affaires russe United Financial Group (UFG) au management de cette dernière. Cette décision, précise-t-il, lui permettra de conduire le développement de ses activités en Russie dans le cadre de son organisation mondiale par métier. Paribas s'appuiera pour cela sur le bureau de représentation qu'il a ouvert en 1966.

La figure B.1 présente l'arbre syntaxique. Les figures B.2, B.3 et B.4 présentent les dépendances syntaxiques extraites pour les expressions apparaissant sous chacun des trois nœuds `ST` de l'arbre de la figure B.1. Sous chaque ensemble de dépendances est reproduite la phrase considérée.

Cet exemple nous donne l'occasion de signaler le point faible de l'analyseur syntaxique : le rattachement des syntagmes prépositionnels. On observe ainsi une erreur d'analyse sur la figure B.2, où le syntagme *au management* est rattaché à *United Financial Group* (l. 11) et une autre erreur sur la figure B.4 où le syntagme *sur le bureau* est rattaché à *cela* (l. 8).

Par ailleurs, on observe que en l'état actuel du système d'analyse syntaxique, des ambiguïtés peuvent apparaître en sortie. Ainsi, sur la figure B.3, le syntagme *dans le cadre de son organisation* est rattaché à la fois à *en Russie* (l. 8) et à *ses activités*, cela étant interprété comme signifiant que ce syntagme est rattaché soit à l'un, soit à l'autre. Il en est de même du syntagme *en Russie* (l. 9 et 11).

Notons que l'analyseur syntaxique était en cours de développement quand nous avons commencé notre travail sur les expressions pronominales et que nous avons dû « figer » une version non définitive pour implanter notre système. Cela explique les erreurs que nous venons d'évoquer.

Signalons pour finir que dans la deuxième phrase, une relation `subj` entre *lui* et *conduire* aurait due être identifiée.

```
1>GROUPE{
  ST{SC{NP{DET{Le},NOUN{groupe}},
        NP{NOUN{Paribas}},
        FV{VERB{va}}},
     IV{VERB{céder}},
     NP{DET{la},NOUN{participation}},
     PP{PREP{de},NP{NUM{25},NOUN{%}}},
     SC{BG{PRON{qu'}},NP{PRON{il}},FV{VERB{détient}}},
     PP{PREP{dans},NP{DET{la},NOUN{banque}}},
     PP{PREP{d'},NP{NOUN{affaires}}},
     AP{ADJ{russe}},
     NP{NOUN{United Financial Group}},
     INS{PUNCT{(},NP{NOUN{UFG}},PUNCT{)}},
     PP{PREP{au},NP{NOUN{management}}},
     PP{PREP{de},NP{DET{cette},NOUN{dernière}}},
     SENT{.}},
  ST{SC{NP{DET{Cette},NOUN{décision}},
     PUNCT{,},
     SC{FV{VERB{précise}}},
     NP{PRON{-t-il}},
     PUNCT{,},
     FV{PRON{lui},VERB{permettra}}},
     IV{PREP{de},VERB{conduire}},
     NP{DET{le},NOUN{développement}},
     PP{PREP{de},NP{DET{ses},NOUN{activités}}},
     PP{PREP{en},NP{NOUN{Russie}}},
     PP{PREP{dans le cadre de},NP{DET{son},NOUN{organisation}}},
     AP{ADJ{mondiale}},
     PP{PREP{par},NP{NOUN{métier}}},
     SENT{.}},
  ST{SC{NP{NOUN{Paribas}},FV{PRON{s'},VERB{appuiera}}},
     PP{PREP{pour},NP{PRON{cela}}},
     PP{PREP{sur},NP{DET{le},NOUN{bureau}}},
     PP{PREP{de},NP{NOUN{représentation}}},
     SC{BG{PRON{qu'}},NP{PRON{il}},FV{VERB{a},VERB{ouvert}}},
     PP{PREP{en},NP{NOUN{NUM{1966}}}},
     SENT{.}}}
```

Fig. B.1 – Arbre syntaxique

```
1  SUBJ(va,groupe)
2  SUBJ(détient,il)
3  SUBJ(céder,groupe)
4  VARG(céder,participation)
5  VARG(va,céder)
6  VARG(détient,qu')
7  VARG(détient,participation)
8  ANTEC(participation,qu')
9  VMOD(détient,dans,banque)
10 NMOD(banque,russe)
11 NMOD(United Financial Group,au,management)
12 NMOD(banque,d',affaires)
13 NARG(participation,de,%)
14 NARG(management,de,dernière)
15 NN(groupe,Paribas)
16 NN(banque,United Financial Group)
17 PREPOBJ(de,%)
18 PREPOBJ(dans,banque)
19 PREPOBJ(d',affaires)
20 PREPOBJ(au,management)
21 PREPOBJ(de,dernière)
22 DETERM(Le,groupe)
23 DETERM(la,participation)
24 DETERM(la,banque)
25 DETERM(cette,dernière)
26 DETERM(au,management)
27 DETERM(25,%)
28 CONNECT(détient,qu')
29
30 Le groupe Paribas va céder la participation de 25%
31 qu'il détient dans la banque d'affaires russe
32 United Financial Group (UFG) au management
33 de cette dernière.
```

Fig. B.2 – Dépendances syntaxiques (1)

```
1  SUBJ(précise,-t-il)
2  SUBJ(permettra,décision)
3  VARG(permettra,lui)
4  VARG(conduire,développement)
5  VARG(permettra,conduire)
6  NMOD(organisation,mondiale)
7  NMOD(activités,en,Russie)
8  NMOD(Russie,dans le cadre de,organisation)
9  NMOD(développement,en,Russie)
10 NMOD(activités,dans le cadre de,organisation)
11 NARG(développement,de,activités)
12 NARG(organisation,par,métier)
13 PREPOBJ(de,activités)
14 PREPOBJ(en,Russie)
15 PREPOBJ(dans le cadre de,organisation)
16 PREPOBJ(par,métier)
17 PREPOBJ(de,conduire)
18 DETERM(Cette,décision)
19 DETERM(le,développement)
20 DETERM(ses,activités)
21 DETERM(son,organisation)
22
23 Cette décision, précise-t-il, lui permettra de conduire
24 le développement de ses activités en Russie
25 dans le cadre de son organisation mondiale par métier.
```

FIG. B.3 – Dépendances syntaxiques (2)

```
SUBJ(appuiera,Paribas)
SUBJ(ouvert,il)
VARG(ouvert,qu')
VARG(ouvert,bureau)
ANTEC(bureau,qu')
VMOD(appuiera,pour,cela)
VMOD(ouvert,en,1966)
NMOD(cela,sur,bureau)
NMOD(bureau,de,représentation)
PREPOBJ(pour,cela)
PREPOBJ(sur,bureau)
PREPOBJ(de,représentation)
PREPOBJ(en,1966)
DETERM(le,bureau)
REFLEX(appuiera,s')
AUXIL(ouvert,a)
CONNECT(ouvert,qu')

Paribas s'appuiera pour cela sur le bureau de représentation
qu'il a ouvert en 1966.
```

Fig. B.4 – Dépendances syntaxiques (3)