Insertions et interprétation des expressions pronominales

François Trouilleux

TALN 2002, Nancy, 24–27 juin 2002

Insertions et interprétation des expressions pronominales

François Trouilleux
GRIL - Université Blaise-Pascal, Clermont II
34, avenue Carnot
63000 Clermont-Ferrand
et
Xerox Research Center Europe
6, chemin de Maupertuis
38240 Meylan
Francois.Trouilleux@xrce.xerox.com

Mots-clefs – Keywords

résolution de l'anaphore, pronom, insertion, apposition, théorie des veines
anaphora resolution, pronoun, insertion, apposition, veins theory

Résumé - Abstract

Dans le contexte d'un système d'interprétation automatique des expressions pronominales en français, nous proposons une contrainte permettant de réduire le nombre d'antécédents potentiels pour un pronom: une expression qui figure dans une insertion ne peut être antécédent d'une expression pronominale qui figure en dehors de cette insertion. La contrainte proposée peut être vue comme un cas particulier des contraintes formulées à un niveau plus général par la théorie des veines. Cependant, les insertions sont ici définies syntaxiquement, ce qui rend notre hypothèse effectivement implantable. L'évaluation sur corpus de notre contrainte sur les insertions donne un taux de succès supérieur à 98 %.

In the context of a pronoun resolution system for French, we propose a constraint to reduce the number of possible antecedents for a pronoun: an expression in an insertion cannot be the antecedent of a pronominal expression which is outside this insertion. The proposed constraint can be seen as a special case of the constraints formulated at a higher level in veins theory. However, we define insertions syntactically, which allows the effective implementation of our hypothesis. Evaluation of the proposed constraint on corpus results in a success rate greater than 98 %.

Le but du présent article est de présenter l'évaluation sur corpus d'une contrainte sur l'interprétation des expressions pronominales, contrainte que nous appelons « contrainte sur les insertions ». L'idée est simple et ne choque pas l'intuition; étant donné le texte suivant,

(1) 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.

elle consiste à dire qu'une expression pronominale ne peut pas avoir pour antécédent une expression qui figure dans l'un ou l'autre des deux segments en *italiques*, ces deux appositions étant considérées comme des « insertions », c'est-à-dire des segments de texte qui ne font qu'apporter des précisions non essentielles par rapport au discours principal.

La contrainte que nous proposons constitue une hypothèse qu'il importe de tester, ce que nous avons fait dans le cadre général de l'implantation d'un système d'interprétation automatique de certaines expressions pronominales en français (Trouilleux, 2001). Le première partie de l'article décrit le contexte d'implantation de la contrainte que nous proposons: un algorithme classique d'interprétation des pronoms en trois étapes principales. Nous définissons ensuite de manière précise la contrainte sur les insertions. Les résultats de l'évaluation sont donnés dans la troisième partie. Enfin, nous terminons l'article par une mise en perspective de notre contrainte et de notre approche avec la théorie des veines (Cristea *et al.*, 1998), qui vise à formuler, à un niveau plus général, des contraintes similaires à la nôtre.

1 Contexte

Le contexte dans lequel nous avons été amené à formuler la contrainte définie ci-après est celui d'un système d'interprétation automatique de certaines expressions pronominales en français (Trouilleux, 2001). Le système défini vise à donner une interprétation unique pour chaque expression pronominale, mais cela sans avoir accès à toute l'information qui semble nécessaire pour interpréter ces expressions [1].

Les systèmes d'interprétation automatique de pronoms implantés faisant usage d'une information incomplète, leurs concepteurs cherchent en général à hiérarchiser l'information utilisée selon son degré de fiabilité, en particulier en distinguant des « contraintes » et des « préférences ». Ainsi, la plupart des systèmes d'interprétation des pronoms implantés à ce jour [2] mettent en œuvre une stratégie de résolution en trois étapes:

1. constitution d'un ensemble d'antécédents potentiels pour chaque expression pronominale,
2. application de contraintes éliminant certains antécédents,
3. détermination de l'antécédent le plus probable par application de préférences.

où les deuxième et troisième étapes se distinguent en partie par le fait que les contraintes ont une validité quasi absolue alors que les préférences ont une validité moindre, clairement statistique [3].

[1] On pense en général que le processus d'interprétation fait appel à des informations de nature sémantique ou pragmatique, mais on ne sait pas bien exprimer celles-ci à l'heure actuelle en traitement automatique.

[2] Voir, par exemple, (Lappin & Leass, 1994), (Mitkov, 1998), (Palomar *et al.*, 2001) et notre propre système (Trouilleux, 2001).

[3] Voir dans (Palomar *et al.*, 2001, § 3, p. 547) une discussion de cette distinction.

En pratique, les contraintes implantées dans les systèmes d'interprétation de pronoms existants sont les contraintes d'accord en genre et nombre et les contraintes syntaxiques [4], ces dernières étant celles qui sont décrites traditionnellement en termes de « c-commande » et/ou de « liage ».

La contrainte que nous proposons dans le présent article s'ajoute à ces deux types de contraintes couramment utilisées. Nous y faisons référence comme une contrainte parce que nous faisons l'hypothèse qu'elle a une validité quasi-absolue lorsqu'évaluée sur corpus.

2 Définitions

Dans un premier temps, nous donnerons une définition « générale et préliminaire » des insertions, puis une seconde définition, plus restrictive, des insertions telles que le système que nous avons défini est susceptible de les identifier. L'hypothèse que nous formulerons ne met en jeu que cette seconde notion d'insertion et sera donc évaluée au regard de la seconde définition.

2.1 Définition générale et préliminaire des 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, une insertion est un segment de texte (i) qui peut être supprimé sans nuire à la correction grammaticale de la phrase, ni à la compréhension globale du discours, et (ii) qui apporte une précision par rapport au discours principal constitué par la phrase sans l'insertion. En outre, les insertions sont des segments de texte marqués comme tels par l'auteur du texte, au moyen de symboles de ponctuation, typiquement des parenthèses ou des virgules. Une insertion est un segment de texte qui, outre les deux propriétés déjà évoquées, (iii) est délimité à gauche et à droite par le début ou la fin de la phrase ou par un symbole de ponctuation.

2.2 Définition restrictive des insertions

En pratique, l'hypothèse que nous formulerons par la suite portera sur *certaines* insertions seulement, que nous caractérisons plus précisément ici.

Notre système est susceptible d'identifier deux types d'insertions: les insertions entre parenthèses, crochets (« [ ] »), accolades (« { } ») ou tirets, d'une part, et des insertions mettant en jeu une virgule à gauche et/ou à droite, d'autre part. Ce sont ces dernières — dites « insertions entre virgules » — qui sont restreintes à quelques contextes particuliers, en l'occurrence les quatre contextes suivants:

1. insertion entre le sujet X d'un verbe fléchi Y et le verbe Y, sauf si X est un pronom relatif. L'insertion est délimitée à gauche par une virgule et à droite par une virgule ou une insertion entre parenthèses (voir l'exemple 2);

[4] Voir, par exemple, le « Syntactic filter on Pronoun-NP coreference », qui inclut l'accord en genre et nombre, de (Lappin & Leass, 1994, p. 537) ou les paragraphes « Morphological agreement » et « Syntactic conditions on NP-Pronoun noncoreference » dans (Palomar *et al.*, 2001, p. 549).

2. insertion entre un verbe fléchi, infinitif ou participe présent et son complément d'objet direct (voir l'exemple 3). Cette règle est restreinte aux compléments d'objets nominaux [5];

3. appositions à droite (voir l'exemple 4);

4. appositions à gauche, à la condition qu'elles portent sur le sujet (voir l'exemple 5 et, *a contrario*, l'exemple 6).

Dans les exemples suivants, sont indiqués en *italiques* les segments de texte considérés comme des insertions.

(2) 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.

(3) Présentant 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 à « mettre fin à l'exception française ».

(4) 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 ».

(5) *Fort de son succès dans la privatisation du CIC,* le gouvernement garde l'ambition de boucler rapidement les autres dossiers financiers aujourd'hui sur la rampe de lancement.

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 restrictive, 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.

(6) Seul regret pour Swiss Life, le CCF*, dont il est actionnaire,* n'a pas été retenu pour le CIC*, partenaire de bancassurance du GAN*.

Dans la mesure où *Swiss Life* est ici antécédent de *il*, cet exemple illustre l'intérêt de notre définition restrictive des insertions. C'est l'observation de tels exemples, où l'hypothèse que nous formulons se révèle invalide si on utilise notre première définition générale des insertions, qui a motivé notre définition restrictive.

Pour terminer, on notera que la définition restrictive des insertions donnée ici ne fait appel qu'à l'information fournie par l'analyseur syntaxique et la présence de symboles de ponctuation. Dans ce sens, cette définition est donc essentiellement *syntaxique*. Nous insistons également sur le fait que l'hypothèse que nous proposons sera évaluée au regard de la définition restrictive seulement, sans préjuger d'un éventuel élargissement de la notion d'insertion qui ferait appel à des notions d'ordre sémantique ou pragmatique.

[5] Par opposition aux compléments propositionnels ou verbaux.

2.3 Hypothèse

La contrainte que nous formulons est la suivante: *une expression pronominale qui se trouve en dehors d'une insertion* i *ne peut avoir pour antécédent une expression figurant dans* i. Le terme *insertion* est employé ici au sens restrictif de la définition donnée au paragraphe 2.2

L'exemple suivant illustre l'application de notre hypothèse:

(7) Swiss Life, beaucoup moins puissant avec un chiffre d'affaires équivalent à celui du GAN, bénéficie du soutien de son actionnaire à 25 %, la United Bank of Switzerland.

Pour cette phrase, notre système d'interprétation des expressions pronominales identifie dans un premier temps comme antécédents potentiels pour le déterminant possessif *son* les expressions: *Swiss Life, un chiffre d'affaires équivalent à celui du GAN, celui du GAN, [le] GAN*. Les trois dernières de ces expressions se trouvant dans une insertion, notre hypothèse est qu'aucune de ces expressions ne peut être antécédent de *son*, qui lui est en dehors de cette insertion. Notre hypothèse est ici confirmée puisque *son* a pour antécédent l'expression *Swiss Life*.

Cet exemple illustre l'intérêt potentiel de la contrainte proposée: l'ensemble des antécédents potentiels de *son* est ici réduit de quatre à un élément, celui-ci étant l'antécédent correct.

Par ailleurs, cet exemple nous invite à préciser ce que nous entendons par le fait qu'une expression pronominale *ne puisse pas* avoir pour antécédent telle ou telle expression. La contrainte que nous formulons ici n'est pas une contrainte sur la structure syntaxique telle que celles qui sont formulées en termes de c-commande ou de liage, mais plutôt une contrainte sur l'usage effectif des expressions pronominales en corpus. Ainsi, dans l'exemple 7, si on se place en terme de c-commande ou de liage, le possessif *son* peut être coréférent avec *le GAN*, mais si on se place en termes d'usage, le lien de coréférence entre ces deux expressions semble exclu.

3 Évaluation

Nous avons évalué notre contrainte sur les insertions dans le contexte d'un système d'interprétation automatique des expressions pronominales en français qui suit globalement la stratégie exposée dans la première partie. Nous présentons successivement les données de l'évaluation puis les résultats obtenus.

3.1 Données

Corpus. Le corpus utilisé est un ensemble d'articles du journal La Tribune, du domaine de la finance [6]. Il est divisé en deux parties: un corpus d'étude (C1), que nous avons observé pour définir notre système et le tester durant le développement, et un corpus d'évaluation (C2), inconnu durant le développement du système et avec lequel nous en avons évalué la version finale. Les corpus C1 et C2 contiennent respectivement environ $23\ 300$ et $18\ 600$ mots, soit un total d'environ $41\ 900$ mots pour l'ensemble du corpus (noté C).

[6] Tous les exemples donnés dans l'article proviennent de ce corpus.

Expressions pronominales visées. Le système vise à spécifier l'interprétation des expressions pronominales de troisième personne suivantes: pronoms personnels sujet (*il, ils, elle, elles*), accusatifs (*l', le, la, les*) et datifs (*lui, leur*), déterminants possessifs (*son, sa, ses, leur, leurs*) et pronoms disjoints (*lui, elle, eux, elles*), avec la restriction que ces expressions renvoient à un syntagme nominal sans coordination et en excluant les formes des pronoms disjoints composées avec *même* (p. ex. *lui-même*), que nous assimilons à des pronoms réfléchis.

Le nombre d'expressions anaphoriques correspondant à ces critères est de 388 dans C1 et de 360 dans C2, soit un total de 748 expressions anaphoriques visées.

Critère et mesure d'évaluation. En sortie de la première étape d'analyse, on a pour chaque expression pronominale e_i appartenant à l'ensemble des expressions visées un ensemble d'antécédents possibles $A(e_i)$. Cet ensemble est constitué d'expressions du contexte proche de e_i. Nous adoptons la convention suivante: le contexte est déterminé par les règles de notre système définissant la première étape du processus d'analyse, si celles-ci permettent d'identifier au moins un antécédent correct pour e_i, sinon en élargissant le contexte à la phrase qui précède le contexte caractérisé par les règles (autrement dit, on fait abstraction des erreurs du système).

Pour notre corpus, ces critères suffisent à identifier pour toute expression e_i, un ensemble $A(e_i)$ qui contient au moins un antécédent correct pour e_i, un antécédent correct étant une expression, pronominale ou non, avec laquelle e_i est coréférente. Nous renvoyons le lecteur à (Trouilleux, 2001) pour une description précise des règles définissant la première étape d'analyse [7].

L'application de la contrainte sur les insertions pour une expression pronominale e_i résulte en un nouvel ensemble d'antécédents possibles $A'(e_i)$.

On juge que la contrainte est *opérante* pour une expression pronominale e_i si $A'(e_i)$ est strictement inclus dans $A(e_i)$ (autrement dit si elle élimine au moins une expression de l'ensemble $A(e_i)$). Soit O le nombre d'expressions pronominales pour lesquelles la contrainte est opérante.

On juge que l'application de la contrainte est *correcte* pour une expression pronominale e_i si elle est opérante et si $A'(e_i)$ contient au moins une expression avec laquelle e_i est coréférente. Soit C le nombre d'expressions pronominales pour lesquelles la contrainte est correcte.

Étant donné les valeurs O et C pour un corpus C*i* donné, on calcule le taux de succès S de la contrainte sur les insertions par la formule $S_{\text{C}i} = C/O$.

3.2 Résultats

Taux de succès. Les taux de succès de la contrainte sur les insertions sont les suivants:

$$S_{\text{C1}} = 79/80 = 98,75\ \% \qquad S_{\text{C2}} = 57/58 = 98,28\ \% \qquad S_{\text{C}} = 136/138 = 98,55\ \%$$

Quelques points sont à noter. En premier lieu, signalons que notre contrainte s'applique pour environ 18 % (138/748) des expressions anaphoriques visées sur l'ensemble du corpus.

Par ailleurs, le ou les antécédents éliminés figurent dans une insertion entre parenthèses dans environ 20 % des cas où la contrainte est opérante. Les deux erreurs relevées concernant des

[7] Signalons que, de manière très générale, le système identifie les contextes de cataphore, ou sinon sélectionne comme antécédents possibles les expressions qui précèdent e_i dans la même phrase ou la phrase précédente, avec une tendance à limiter le contexte à la même phrase.

insertions entre virgules, si on calcule le taux de succès global en ne prenant en compte que ce dernier type d'insertion, on obtient une valeur légèrement inférieure de 98,18 % (108/110) pour le taux de succès global S_{C}.

Enfin, il faut également noter que la contrainte sur les insertions est ici évaluée sans préjuger des éventuelles autres contraintes, en particulier les contraintes d'accord, qui pourraient également s'appliquer pour éliminer des antécédents potentiels éliminés par la contrainte sur les insertions.

Contribution relative de la contrainte sur les insertions. Pour tenter de cerner l'intérêt de la contrainte évaluée ici, on peut mesurer sa contribution à la réduction du nombre d'antécédents possibles pour l'ensemble des expressions pronominales visées. C'est ce que nous faisons ici, en comparant les résultats obtenus avec ceux qui sont produits par les contraintes d'accord [8].

L'évaluation est ici effectuée sur l'ensemble des expressions pronominales visées pour lesquelles le système identifie au moins un antécédent correct en sortie de l'étape 1. Soit N_e l'ensemble de ces expressions. Soit $R(e_i)$ le nombre total de « référents » possibles [9] pour chacune des expressions pronominales $e_i \in N_e$. La division suivante donne R_m, le nombre moyen de référents possibles par expression pronominale:

$$R_m = \frac{\Sigma_{e_i \in N_e} R(e_i)}{|N_e|}$$

En sortie de l'étape 1, on a donc une valeur pour $|N_e|$ (cardinalité de N_e) et une valeur R_m. Si on applique une contrainte ou un ensemble de contraintes sur la sortie de l'étape 1, on obtient de nouvelles valeurs $|N_e|$ et R_m. On évalue la contribution de la contrainte ou de l'ensemble de contraintes en mesurant le pourcentage de réduction (%R) de la valeur R_m, sachant qu'en pratique, la valeur $|N_e|$ varie peu (il y a peu d'erreurs).

Le tableau de la figure 1 donne la contribution de trois contraintes différentes, appliquées chacune sur la sortie de l'étape 1 de notre système d'interprétation des expressions pronominales [10]. On constate que chacune des contraintes produit un nombre d'erreurs similaire (colonnes $|N_e|$). La contrainte d'accord en nombre est celle qui permet le mieux de réduire le nombre de référents possibles. La moindre contribution de la contrainte d'accord en genre est due principalement au fait que bon nombre d'expressions pronominales (les déterminants possessifs, les clitiques *lui, leur, l'* et *les*) ne contiennent aucune indication quant au genre de leur antécédent. Enfin, pour ce qui concerne la contrainte sur les insertions, on constate qu'elle permet une réduction non négligeable, et assez fiable au vu des taux de succès obtenus, du nombre de référents possibles pour les expressions pronominales considérées.

Les résultats obtenus ici montrent l'intérêt de la contrainte proposée. Il convient cependant de les relativiser: il est probable que la contribution de la contrainte sur les insertions sera minime lorsqu'elle sera appliquée à des corpus de domaines différents, par exemple, des manuels techniques. En effet, si les insertions sont en règle générale assez nombreuses dans les articles de presse, elles sont probablement assez rares dans d'autres domaines.

[8] Nous ne prenons pas en compte les contraintes syntaxiques, celles-ci étant implantées dans notre système d'une façon qu'il serait trop long de décrire.

[9] Le nombre de référents possibles pour une expression pronominale est déterminé comme suit: chaque ensemble $A(e_i)$ est partitionné en n classes d'équivalence caractérisées par la relation de coréférence et $R(e_i) = n$.

[10] Les chiffres donnés prennent en compte toute erreur du système, ce qui explique qu'on recense deux erreurs dues à la contrainte sur les insertions pour C2 (l'une de ces erreurs n'est donc pas due à l'hypothèse elle-même).

	corpus d'étude (C1)			*corpus d'évaluation* (C2)		
	$\|N_e\|$	R_m	%R	$\|N_e\|$	R_m	%R
sortie étape 1	374	4,101		329	4,106	
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

Figure 1: Contribution de différentes contraintes.

3.3 Contre-exemple(s)

Les deux contre-exemples que nous observons sont les suivants [11]:

(8) 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.

(9) La bonne tenue de l'activité 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.

L'exemple 8 est clairement un contre-exemple, pour lequel nous n'avons pas d'explication. On remarquera que le syntagme déterminé par *son* et l'insertion qui contient son antécédent font ici partie d'un même syntagme nominal, dont le noyau est *stratégie*. Nous envisageons d'étudier ce type de configuration, sachant que l'algorithme de (Hobbs, 1976) exprime dans ce cas de figure une préférence — ici correcte — pour un antécédent figurant dans le même syntagme.

L'exemple 9 n'est pas véritablement un contre-exemple à notre hypothèse. Le problème vient ici plutôt de la détermination de l'ensemble d'antécédents initial [12]: notre système identifie comme antécédents potentiels seulement les expressions qui précèdent *leurs* dans la même phrase. Dans ce contexte, *leur*, éliminé par la contrainte, est le seul antécédent correct. Si on élargit le contexte aux expressions de la phrase précédente, ce qui semble une stratégie raisonnable, on retrouve cependant un autre antécédent correct: *les banques françaises*.

4 Discussion

La contrainte sur les insertions que nous proposons n'est pas sans rapport avec la « théorie des veines » (Cristea *et al.*, 1998; Ide & Cristea, 2000), qui vise à formuler, mais à un niveau plus général, une contrainte similaire sur l'interprétation des expressions anaphoriques.

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 & Thompson, 1988), qui représente

[11] L'insertion est en *italiques*, l'antécédent rejeté par la contrainte entre crochets et l'expression anaphorique en petites capitales.

[12] Voir le critère d'évaluation exposé plus haut (§ 3.1), en particulier la convention sur la détermination du contexte et la tendance du système, évoquée dans la note 7, à sous-estimer ce contexte.

A.	[A]	Michael D. Casey, a top Johnson & Johnson manager, moved to Genetic Therapy Inc.,a small biotechnology concern here,
B.	[AB]	to become its president and chief operating officer.
C.	[AC]	Mr. Casey, 46 years old, was president of J&J's McNeil Pharmaceutical subsidiary,
D.	[ACD]	which was merged with another J&J unit, Ortho Pharmaceutical Corp., this year in a cost-cutting move.
E.	[AE]	Mr. Casey succeeds M. James Barrett, 50, as president of Genetic Therapy.
F.	[AEF]	Mr. Barrett remains chief executive officer
G.	[AEFG]	and becomes chairman.
H.	[AH]	Mr. Casey said
I.	[AHI]	he made the move to the smaller company
J.	[AIJ]	because he saw health care moving toward technologies like the company's gene therapy products.

Figure 2: Unités de discours et domaines d'accessibilité.

la structure d'un texte par un arbre définissant les relations entre les différentes parties du texte, relations caractérisées le plus souvent comme des relations entre des *noyaux* et des *satellites*.

À partir de cette structure, (Cristea *et al.*, 1998) définissent la notion de « veine », puis celle de « domaine d'accessibilité » [13], à partir de laquelle est formulée la conjecture suivante: « les références à partir d'une unité de discours [14] donnée ne sont possibles que dans son domaine d'accessibilité. »

La figure 2 reproduit un texte donné en exemple dans (Ide & Cristea, 2000). Chaque ligne constitue une unité de discours, identifiée par une lettre en début de ligne, lettre suivie d'une séquence de lettres entre crochets qui spécifie le domaine d'accessibilité de l'unité en question. Par exemple, le domaine d'accessibilité de l'unité I est constitué des segments A, H et I et, en effet, les trois expressions référentielles du segment I (*he*, *the move to the smaller company* et *the smaller company*) sont bien coréférentes avec des expressions figurant dans A ou H.

(Cristea *et al.*, 1998, 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* [15]. 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. »

Dans la mesure où, dans une paire <noyau,satellite>, le noyau est supposé exprimer une information plus importante [16], et où, dans ce sens, les insertions peuvent être vues comme un cas particulier de satellites [17], on sera tenté de penser que la contrainte sur les insertions décrite ici n'est qu'un cas particulier de celle qui est formulée par la théorie des veines, mais une différence importante est à noter.

[13] Pour une définition de ces notions et de l'algorithme qui permet de passer des structures rhétoriques aux veines, voir (Cristea *et al.*, 1998) ou (Ide & Cristea, 2000).

[14] C'est-à-dire une partie minimale, non décomposée, de texte.

[15] C'est nous qui soulignons.

[16] Voir (Mann & Thompson, 1988, § 10.1).

[17] (Marcu *et al.*, 1999) font usage d'une relation rhétorique appelée « apposition ».

Pour que les domaines d'accessibilité de la théorie des veines soient un moyen de contraindre la résolution de l'anaphore, il faudrait que les structures rhétoriques à partir desquelles ces domaines sont définis (via les « veines ») soient identifiables sans interpréter les expressions anaphoriques, or cela est probablement impossible. À titre d'exemple, on voit mal comment, dans le texte de la figure 2, le rattachement de l'unité de discours I au reste du discours pourrait être déterminé sans interpréter les trois expressions anaphoriques que cette unité contient.

Les insertions que nous avons définies, en revanche, sont identifiables sur la base d'informations syntaxiques et sans recourir à l'interprétation des expressions pronominales, ce qui fait que notre contrainte sur les insertions peut effectivement servir à contraindre l'interprétation des expressions pronominales.

5 Conclusion

Nous avons proposé et testé une contrainte sur l'interprétation des expressions pronominales dans les discours textuels: une expression figurant dans une insertion ne peut être l'antécédent d'un pronom qui se trouve en dehors de cette insertion. L'évaluation de cette contrainte, telle qu'effectivement implantée dans un système d'interprétation automatique des pronoms, confirme dans une large mesure l'hypothèse proposée. L'étendue relativement réduite du corpus utilisé, ainsi qu'un contre-exemple observé dans ce corpus, suggèrent cependant qu'une analyse plus poussée devra être menée à bien dans le futur.

Références

CRISTEA D., IDE N. & ROMARY L. (1998). Veins theory: A model of global discourse cohesion and coherence. In *Proceedings of COLING-ACL'98*, p. 281–285, Montréal, Canada.

HOBBS J. (1976). Resolving pronoun references. *Lingua*, **44**, 311–338.

IDE N. & CRISTEA D. (2000). A hierarchical account of referential accessibility. In *Proceedings of ACL-2000*, p. 416–424, Hong Kong.

LAPPIN S. & LEASS H. J. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*, **20**(4), 535–561.

MANN W. C. & THOMPSON S. A. (1988). Rhethorical structure theory: Toward a functional theory of text organization. *Text*, **8**(3), 243–281.

MARCU D., AMORRORTU E. & ROMERA M. (1999). 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.

MITKOV R. (1998). Robust pronoun resolution with limited knowledge. In *Proceedings of COLING-ACL'98*, p. 869–875, Montréal, Canada.

PALOMAR M., FERRÁNDEZ A., MORENO L., MARTÍNEZ-BARCO P., PERAL J., SAIZ-NOEDA M. & MUÑOZ R. (2001). An algorithm for anaphora resolution in Spanish texts. *Computational Linguistics*, **27**(4), 545–567.

TROUILLEUX F. (2001). *Identification des reprises et interprétation automatique des expressions pronominales dans des textes en français*. Thèse de doctorat, GRIL, Université Blaise-Pascal, Clermont-Ferrand.