Lexique et classement en parties du discours dans Orfeo

Henri-José Deulofeu, André Valli

Lexique et classement en parties du discours dans Orfeo

Pos tagset and lexicon in Orféo

José Deulofeu

André Valli
TALEP Laboratoire Informatique et Systèmes
Université d'Aix-Marseille

Résumé

L'article présente les principes et les critères qui ont présidé à l'élaboration de la table des parties du discours et à l'organisation du lexique correspondante, mis en œuvre dans l'analyse syntaxique automatique du corpus Orféo . La comparaison est établie avec le Lexique des Formes Fléchies du Français (Lefff) utilisé dans d'autres outils de traitement automatique du langage. Les enjeux linguistiques et informatiques sont abordés. Un développement particulier est consacré au traitement des locutions ou expressions multi-mots. Des perspectives d'amélioration sont envisagées.

Mots-clés

Parties du discours, lexique, locutions, annotation

Abstract

The paper discusses the principles and criteria used in elaborating the POS tagset and the structure of the corresponding lexicon at use for the automatic parsing of the Orféo corpus. . This architecture is compared with the current Lexique des Formes Fléchies du Français (Lefff) dictionary, available under Open source license. The linguistic and natural language processing challenges are dwelled on. A specific attention is devoted to the processing of multiword expressions. Some ways of improvement of the system are provided.

Keywords

Parts of speech, multiword expressions, annotation, lexicon

1. Introduction

Les systèmes d'annotation automatique s'appuient généralement sur des lexiques électroniques qui sont dérivés de dictionnaires hérités de la tradition lexicographique. Nous sommes partis du plus courant de ces lexiques, le Lexique des Formes Fléchies du Français (Lefff (Sagot et *al*. 2006), mais nous l'avons modifié pour l'adapter au but propre à Orféo qui est de permettre aux linguistes de réaliser des requêtes fiables de patterns morphosyntaxiques, c'est-à-dire des séquences de mots étiquetées en un ensemble de catégories lexicales ou parties du discours reliées par les relations syntaxiques (Voir l'article de Kahane & *al*. dans ce numéro).

Nous rappelons que le choix d'un ensemble d'étiquettes morphosyntaxiques ne reflète pas nécessairement la meilleure analyse linguistique de la langue traitée. Notamment parce que la description des langues n'étant pas encore achevée et les résultats ne faisant pas l'unanimité dans la communauté, il n'est pas possible de proposer une référence absolue à laquelle on pourrait se référer. Mais cette limitation n'est pas un obstacle, si on rappelle que les dispositifs d'analyse automatique ne produisent pas des analyses, mais sont des outils pour améliorer les analyses existantes. Les systèmes d'étiquettes interviennent à un niveau préalable à l'analyse, celui de l'établissement des faits à analyser. Et l'on sait que l'analyse aboutit souvent à remettre en cause les catégories de classement initiales. La ressource Orféo (Outils pour la Recherche sur le Français Ecrit et Oral) est un dispositif d'aide à l'établissement des faits et non pas un outil d'analyse de ces faits. La pertinence des étiquettes morphosyntaxiques se mesure principalement aux performances de l'outil de requête, plus particulièrement à la capacité de ce système à regrouper de façon claire et rapide des exemples tirés de corpus répondant à un certain schéma morphosyntaxique. Le défi pour le système est de permettre au linguiste de récupérer tous les exemples du corpus utiles pour son analyse. En termes quantitatifs, la mesure qu'il convient d'évaluer est plutôt le « rappel » que la « précision », c'est-à-dire la capacité du système à extraire du corpus le plus possible d'exemples répondant à un certain schéma, même si les résultats des requêtes contiennent des exemples non pertinents qu'il sera facile d'éliminer manuellement. Beaucoup de nos choix ont été guidés par cet objectif.

Comme le jeu d'étiquettes que nous avons choisi pour atteindre cet objectif ne recouvre pas la même organisation du lexique que celle du Lefff, notre premier travail a consisté à expliciter les critères utilisés pour le classement des formes en parties du discours et, par conséquent, pour définir les étiquettes correspondantes (POS). Il est en effet nécessaire que l'utilisateur de l'outil de requêtes puisse savoir à quoi il peut s'attendre lorsqu'il formule une requête libellée mettant en jeu des étiquettes parties de discours. S'il fait par exemple des recherches sur la « subordination », la fiabilité de ces recherches implique notamment qu'il sache exactement quel est le contenu de l'étiquette CSU (conjonction de subordination) de la table des catégories Orféo, élément essentiel de ces requêtes. Or, le contenu de l'étiquette diffère selon les approches : les lexiques apparentés au Lefff conformément à une tradition couramment admise y incluent les « locutions conjonctives » comme *avant que*, *pour que*, etc. D'autres approches, fondées sur des critiques de la tradition, les excluent de la liste des CSU pour en faire des suites régulières PRE + CONJ dont les attestations ne pourront être trouvées que par des requêtes spécifiques. L'utilisateur doit donc être précisément informé de notre choix et des principes sur lesquels repose notre classification des lexèmes en parties du discours.

Il arrivait souvent que plusieurs choix de classement soient possibles. Nous avons choisi le jeu d'étiquettes que nous présentons ici en articulant au mieux possible les trois objectifs suivants :

- appuyer les classements sur des critères linguistiques robustes et cohérents ;
- faciliter le travail du logiciel d'annotation automatique ;
- s'adapter aux connaissances d'un public d'utilisateurs le plus divers possible.

Nous détaillons dans ce qui suit les choix qui ont présidé au classement des items lexicaux en parties du discours et par voie de conséquence la structure générale du lexique qui a servi de base aux annotations morphosyntaxiques manuelles et semi-automatiques du Corpus d'Étude du Français Contemporain issu du projet Orféo. Nous montrerons que les réflexions sur la constitution d'un outil peuvent amener à discuter et à renouveler des questions importantes de description du français. Nous terminerons en évoquant les perspectives d'amélioration de l'outil rendues possibles par sa mise à disposition sur le site www.ortolang.fr, où sont rassemblées de nombreuses ressources sur la langue française. La liste des étiquettes avec leur nom complet est donnée en annexe.

2. CRITERES DE DISTINCTION ENTRE CATEGORIES

2.1. Principes généraux

2.1.1. Catégories ouvertes versus catégories fermées

Nous avons considéré que le problème majeur dans l'élaboration du dictionnaire tenait moins au contenu des catégories « ouvertes » ou lexicales (Nom, Verbe, Adjectif) qu'à celui des catégories « fermées » ou fonctionnelles (PREpositions, CONJonctions, INTerjections, DETerminants, PROformes). Ces dernières interviennent de façon cruciale dans beaucoup de débats syntaxiques et, à la différence des premières, elles ne sont pas distinguées par une morphologie particulière. Nos propositions de modification concernent donc essentiellement le domaine des catégories fonctionnelles. Un cas particulier concerne le traitement de l'ADVerbe dont le statut est problématique y compris au regard de la distinction lexicale-fonctionnelle.

2.1.2. Relation catégorie-fonction

Nous avons cherché à limiter le nombre de cas où un item est classé dans deux catégories différentes. Pour cela, nous avons tiré parti des multiples possibilités d'articuler catégorie et fonction. Nous avons donc traité syntaxiquement des problèmes qui auraient pu aboutir à une double appartenance catégorielle de certains items. Nous avons autorisé certaines catégories à dépendre par un même lien syntaxique de deux gouverneurs différents : ainsi *juste, fort, vrai* sont classés uniquement comme adjectifs. Leur emploi dit adverbial dans *jouer juste, chanter fort, parler vrai* est traité comme une dépendance régulière au verbe. Nous avons posé une relation syntaxique particulière de dépendance au gouverneur nominal SPE (spécifieur), de sorte que nous n'avons pas confondu sous la même étiquette DET une catégorie et une relation syntaxique. La relation SPE, propre aux gouverneurs nominaux, accepte comme dépendants aussi bien des DET que des membres d'autres catégories, par exemple les adjectifs : *différent* est ainsi classé seulement comme adjectif. Dans *différentes solutions peuvent être envisagées, différentes* est un adjectif en fonction de SPE. De même, *quel* est classé seulement comme PRQ (pronom *qu-*) et peut occuper une fonction SPE dans *quel livre voulez-vous*, tandis qu'il sera analysé comme dépendant direct du verbe dans *quel est-il ?*

Selon ce même principe, nous n'avons pas classé des lexèmes tels que *malheur* et *merde* dans deux catégories selon leur fonctionnement comme nom ou

interjection. Nous avons préféré leur conserver la catégorie NOM et traiter syntaxiquement leur emploi comme interjection par une relation DM (marqueur de discours), ce qui nous a encore permis de limiter les cas de double classification. Ces doubles classifications existent évidemment dans le cas de termes aux distributions très différenciées comme *sort*, *plonge* (verbe et nom) etc. Et même dans le domaine des catégories fonctionnelles, le principe n'a pas toujours pu être appliqué. Le détail des décisions sera discuté par la suite.

2.1.3. Séquences régulières de catégories ou expressions composées

Nous avons cherché le plus possible à analyser comme régulières des séquences qui auraient pu être traitées comme des unités composées de plusieurs mots. Nous détaillerons plus loin (section 3.2) les critères utilisés pour classer une séquence comme une catégorie multi-mots. Nous avons traité essentiellement les cas de locutions fonctionnelles prépositionnelles ou conjonctionnelles. Notre stratégie a consisté à limiter le nombre de locutions fonctionnelles (prépositions, conjonctions, déterminants). En effet, les locutions fonctionnelles figurant dans le lexique et qui sont homonymes de séquences régulières empêchent l'analyseur de proposer une analyse différente de séquences comme *bien que* dans :

1. Je sais bien qu'il est venu.
2. Il est venu bien que je le lui aie interdit.

La double analyse serait impossible si la séquence *bien que* était d'emblée classée dans le lexique comme conjonction complexe.

Pour d'autres cas, il suffisait de suivre des analyses linguistiquement justifiées qui proposent de considérer qu'une locution potentielle peut être analysée comme une séquence régulière de catégories. Si on considère, par exemple, qu'une préposition ne peut admettre de complément phrastique, une séquence comme *avant que* ne sera pas considérée comme une suite régulière. Il faut alors l'analyser comme une conjonction complexe *avant que*. C'est la solution couramment choisie dans les lexiques TAL, conformément à la grammaire scolaire.

Mais on peut aussi autoriser les prépositions à gouverner des *que*-Phases, en accord avec les théories syntaxiques récentes qui proposent d'aligner la valence des prépositions sur celle des verbes. Cette option permet de limiter directement le nombre d'expressions complexes et donc d'erreurs d'analyse potentielles. Ainsi la séquence en gras dans *il a été mis en* ***avant que*** *jusqu'ici le Comité avait fait une analyse trop étroite de ses relations avec les médias* peut recevoir une analyse distincte de celle qui lui est attribuée dans *il faut agir* ***avant qu'****il ne soit trop tard*.

2.2. Critères

Des débats théoriques existent qui mettent en doute la possibilité de définir des parties du discours comme des sous ensembles disjoints de lexèmes. Soit que l'on considère, comme Croft (2000), que les catégories doivent être définies pour chaque construction de la langue, soit comme Aarts (2004) qu'il faut une définition vectorielle prenant en compte la grande variété des comportements syntaxiques. Nous avons pris la décision pratique de définir nos catégories comme des ensembles disjoints pour faciliter le travail de requête. Il s'agissait donc d'articuler de façon opératoire des critères relevant de plusieurs domaines :

– internes au lexème : morphologie variable versus invariable ;

– externes : distribution de l'élément en fonction du contexte syntaxique
  - gouverneurs et dépendants possibles (valence passive et active)
  - critères « topologiques » : position préverbale (pour les clitiques)
  - *ad hoc* permettant un repérage intuitivement facile de la catégorie.

C'est par exemple un tel critère *ad hoc* qui nous a amenés à créer une catégorie NUM, définie par extension comme comportant les numéraux cardinaux. La distribution très large et particulière des numéraux cardinaux nous a conduits à en faire une catégorie propre pour éviter les classements multiples et limiter les erreurs d'analyse automatique.

Une fois prise cette décision, il nous fallait hiérarchiser les critères de classification. Nous avons pris comme point de départ le critère interne qui oppose les formes variables et les formes invariables. Ce critère en recoupe un autre plus lexical, celui qui oppose les catégories « ouvertes » ou productives et les catégories « fermées » qui n'accueillent pas librement de nouveaux membres. Ces catégories fermées sont aussi appelées fonctionnelles en ce qu'elles jouent un rôle important dans l'organisation de la structure grammaticale, notamment comme marqueurs de relations syntaxiques.

Les seules catégories qui comportent à la fois des variables et des invariables sont les catégories « fermées » fonctionnelles CLI (clitiques) et les PRQ (interrogatifs relatifs), classes fermées que l'on peut en outre aisément caractériser par une distribution spécifique fondée sur la topologie.

À l'intérieur des grandes catégories définies par l'opposition variable / invariable, plusieurs options de classement sont toujours possibles en articulant sous-catégorisation et topologie. Nous détaillons ci-dessous nos choix. Nous proposerons en conclusion des pistes de recherche qui pourraient nous permettre d'améliorer les performances de l'analyseur.

3. LE CLASSEMENT EN PARTIES DU DISCOURS

Ce classement reflète les consignes qu'ont suivies les annotateurs dans la tâche d'étiquetage du corpus d'entrainement.

3.1. Catégories variables

Pour les catégories « ouvertes » : verbes, noms, adjectifs, nous sommes restés fidèles au lexique du Lefff. Nous avons choisi de ne pas construire de sous catégories en fonction de caractéristiques morphologiques : masculin / féminin, pluriel / singulier.

3.1.1. Verbe

On distingue les étiquettes VRB (verbe fini ou conjugué), VPP (verbe au participe passé), VPA (verbe au participe présent), VIF (verbe à l'infinitif). Certains lexèmes se voient alors attribuer deux POS. La répartition obéit à des critères syntaxiques.

Exemple du classement en VPP ou ADJ pour le lexème *serré* :

– ADJ quand la distribution est celle d'un adjectif :

3. Je veux un café très serré.

– VPP si distribution verbale :

4. Le garagiste a trop serré la vis.

5. La vis qui a été trop serrée par le garagiste n'a pas tenu.

3.1.2. Nom

On ajoute au critère morphologique de variabilité le critère syntactico-sémantique qu'un lexème, par ailleurs classé ADJ, peut être classé comme nom dans des cas où il est tête de SN avec sens indépendant du contexte. On analyse donc *bleu* comme NOM dans :

6. Les Bleus, un bleu (de travail)

et comme ADJECTIF dans :

7. La boule blanche et la bleue (COO DET ADJ)

Mais nous avons considéré également des cas particuliers. En effet, *question, rapport, côté* sont doublement catégorisés. Ils sont analysés comme des NOMs par défaut, mais ils peuvent également être PRE dans :

8. Question poisson, je préfère le thon.

La distribution de *question poisson* est celle d'un groupe prépositionnel, dont *question* est le gouverneur, ce qui justifie le changement de catégorie.

On garde en revanche la catégorie NOM pour le modifieur du gouverneur *paquet* dans *un paquet cadeau, un exemple limite, un livre témoignage*, même s'il présente des caractéristiques distributionnelles d'ADJ comme dans *un exemple très limite*. De même, *limite* reste NOM dans *cette attitude c'est limite une agression.* Dans tous ces cas, la présence d'un modifieur nominal n'entraîne pas un changement de gouverneur.

3.1.3. Adjectif

Deux décisions méritent d'être signalées. Des lexèmes classés traditionnellement comme déterminants ont été reversés dans la catégorie ADJ (voir la section DET). Au contraire, pour les néologismes comme *il est super, trop*, *juste* notre dictionnaire admet le classement comme adjectif à côté de celui comme adverbe suivant en cela le Lefff.

La composition des Catégories variables « fermées » présente des différences plus nombreuses avec le classement du Lefff.

3.1.4. Déterminant

Cette catégorie regroupant traditionnellement articles et déterminants a été restructurée en fonction d'analyses linguistiques récentes et de considérations liées au TAL. Pour les objets « directs » partitifs et indéfinis, comme dans *je lis des livres* ou *je veux de la tarte*, nous avons choisi l'analyse qui, s'appuyant sur l'équivalence *je lis des livres / j'en lis*, conduit à décomposer systématiquement *du, des* en préposition + déterminant (*de + le/ les/la*) dans tous les cas. Cela signifie qu'il n'y a pas dans notre lexique d'article ou de déterminant partitif, ni d'article indéfini pluriel en tant que tel. Pour les autres cas, la fonction SPE est attribuée aux lexèmes qui permettent de faire fonctionner un NOM comme sujet[1]. On classe ainsi dans DET les lexèmes simples qui font fonctionner un N comme sujet et qui ne sont pas combinables entre eux : *le, ce, mon, un, certains, chaque, plusieurs, quelque, zéro, aucun, tout, n'importe quel*.

[1] Pour la définition détaillée des fonctions, voir l'article de Kahane & *al*. dans ce volume.

On classe comme adjectifs faisant fonction de SPE ceux qui peuvent se combiner avec les précédents : *(ces) quelques, (les) différents, (un) autre/même/certain.*

Ces éléments ont par ailleurs un fonctionnement d'adjectif : *c'est certain / différent / tout autre*.

Pour *quel*, qui peut fonctionner à la fois comme SPE (*quel homme*) et comme PRQ interrogatif (*quel est-il*), on choisit un classement unique PRQ. Puisqu'on admet que des ADJ peuvent fonctionner aussi comme SPE, on peut étendre cela aux PRQ.

Enfin, dans le souci de traiter le plus possible les locutions comme des expressions régulières, on ne considère pas comme DET les « déterminants complexes ». *Beaucoup de X, trop de, plein de, la plupart de* sont donc décomposés. Le premier élément fonctionne comme tête avec sa catégorie habituelle : ADV (*trop*, *plus*…) ou PRO (*beaucoup, la plupart*) ou ADJ (*plein*).

Ainsi, syntaxiquement, la fonction SPE concerne tous les mots qui permettent à un SN de fonctionner comme sujet (qu'ils soient ADJ, PRQ ou DET). La position sujet accepte en outre des séquences dont la tête est la PRE *de* (*des passants se hâtaient*) ou encore des adverbes (*trop* de, *plus* de).

3.1.5. Proforme

Les proformes constituent une classe fermée de lexèmes qui peuvent remplir à eux seuls les fonctions de groupes nominaux, adjectivaux ou prépositionnels. Cependant, nous avons suivi le Lefff en classant comme pronoms (PRO) les formes variables à distribution de SN, et en renvoyant les proformes invariables à distribution de groupes prépositionnels (*ici, là, ainsi, autant, n'importe où quelque part*) à la catégorie adverbe. Nous nous en sommes distingués en établissant les étiquettes CLS, CLN[2], PRQ qui rassemblent des lexèmes variables et invariables, des pro-SN (*il, le, qui*) et des pro-SP *(y, en, où*).

Sont classés PRO des items variables qui peuvent à eux seuls remplir les fonctions et occuper les positions d'un groupe nominal : *celui-ci, lui, ce, n'importe qui, qui que ce soit*. *Tel,* qui est en fait un pro-SAdj (*il est tel, tel il s'est présenté devant vous*), est classé PRO. Nous avons créé une étiquette pour les proformes présentant la distribution particulière des clitiques, qu'elles soient variables (*il, ils, lui/ leur*) ou invariables (*y, en*). Avec une étiquette particulière CLS pour le clitique sujet, qui joue un rôle important dans la syntaxe de la langue parlée. Quant à l'étiquette PRQ, celle-ci rassemble les interrogatifs et relatifs, sur la base de leur distribution spécifique en tête de construction. Comme pour les clitiques, on trouve des PRQ variables *qui/quoi* (pro SN) et des PRQ invariables *où, quand* (pro SP). *Qui* et *que* relatifs sujet et objet sont classés comme pronoms et non comme des conjonctions.

3.2. Catégories invariables

3.2.1. Principes

Une possibilité évoquée par Jespersen (1924) aurait été de considérer qu'il n'y avait qu'une catégorie de lexèmes invariables. Avec l'argument que les divers

[2] CLN correspond au *ne* de négation.

contextes syntaxiques couramment utilisés pour établir des catégories ne distingueraient en fait que des sous-catégories comme dans le cas des verbes ou des noms.

Mais, outre la trop grande généralité des requêtes formulées à partir cette catégorie unique, deux arguments structuraux vont contre cette position radicale. Certains de ces items n'ont en fait pas de sous-catégorisation propre : ceux qui fonctionnent comme des COO (conjonctions de coordination). Dans les coordinations, la sélection de leur « complément » est en fait déterminée par le gouverneur de la coordination. Plus généralement, seuls ceux qui sont classés traditionnellement comme PRE et certains adverbes ont un éventail de sous-catégorisations diversifiées comparables à celles du verbe, du nom et de l'adjectif. Les conjonctions prototypiques, par exemple, *que*, ou *parce que*, ne se combinent librement qu'avec des constructions verbales finies. Une propriété qui les distingue radicalement des verbes. Il faut prendre aussi en compte qu'il existe des items, les interjections, qui ne sont pas intégrables dans la structure grammaticale de l'énoncé et pour lesquels il n'y a donc pas de contextes de sous-catégorisation.

Ces considérations nous ont conduits à faire une première partition des invariables sur la base de propriétés topologiques. On peut en effet distinguer les invariables qui figurent systématiquement en tête de construction (PRE, CSU, COO) de ceux dont la distribution est plus diverse (ADV, INT).

Pour les invariables en « tête de construction », on maintient les distinctions entre PRE, CSU et COO. Le classement s'écarte cependant de la tradition sur les points suivants. La distinction entre PRE, CSU et COO repose principalement sur la propriété que les PRE ne peuvent construire des VRB à la différence des CSU.

9. *Pour il aille à Toulon
10. (qu'/ comme / si / quand) il va à Toulon

Il est en effet possible d'aligner les contextes de sous-catégorisation des verbes avec ceux des prépositions, ce qui nous a notamment permis de considérer comme régulières les séquences correspondant aux conjonctions composées de la tradition *pour que, après que…* décomposées en PRE + CSU. En revanche, en français, les verbes ne sous-catégorisent pas des phrases finies. Il est donc cohérent de regrouper les prépositions d'un côté et les conjonctions de l'autre.

Les COO peuvent alors être distinguées des CSU par leur absence de valence active. Pratiquement, cette propriété un peu abstraite peut être recoupée avec la propriété classique de possibilité de reprise par la séquence *et que*.

Nous verrons que dans le détail la partition n'est pas parfaite : nous avons été amenés à admettre de multiples catégorisations, par exemple pour *comme* (PRE, CSU) ou *ainsi que* (COO, CSU).

Pour les autres invariables, si la définition des interjections (étiquette INT) par leur impossibilité d'intégration à la structure grammaticale est opératoire, on rencontre plus de difficultés pour définir la classe des adverbes, comme il sera discuté plus loin.

3.2.2. Détail des parties du discours invariables

3.2.2.1. Préposition (PRE)

La propriété caractéristique de gouverner une CSU doit être précisée, car elle est partagée par les adverbes en *-ment* dans :

Commenté [relecteur1]: :

11. Heureusement qu'il est venu.

Il faut ajouter une caractéristique de valence passive pour distinguer les deux emplois. Les PRE gouvernent les CSU à la fois dans des phrases racines et dans des dépendantes :

12. Avant que je revienne ! / Fais-le avant que je revienne.

Les adverbes ne gouvernent des CSU que dans des phrases racines (principales) :

13. Heureusement qu'il est venu. / *Je pense que heureusement qu'il est venu.

Une autre option aurait été de faire de *heureusement que* une CSU composée sur le modèle de *surtout que, alors que, bien que*. Mais les composés adverbiaux de *que* peuvent fonctionner à la fois en racine ou en dépendante. Nous avons donc préféré ne pas briser cette régularité et faire de l'adverbe un gouverneur, ce qui permet en outre de maintenir un parallélisme intéressant avec l'adjectif correspondant :

14. Il est heureux qu'il soit venu.

3.2.2.2. Conjonction de subordination (CSU)

Notre définition qui permet de regrouper *que, si, comme*, recoupe celle du *complementizer* de la grammaire générative d'inspiration chomskienne. Cependant, dans notre cadre de grammaire en dépendances, la CSU partage avec les adverbes connecteurs *donc, en effet, ensuite* la propriété de se combiner avec une construction verbale finie. C'est la propriété topologique d'être en tête de la construction qui l'en distingue.

Nous avons proposé une seconde caractéristique quelque peu *ad hoc*, le fait de pouvoir être reprise par *et que*, qui permet d'exclure de la catégorie CSU *car* et *or*, pour les reclasser comme COO (conjonction de coordination) bien qu'ils possèdent une valence propre puisque ces items sous catégorisent strictement des constructions verbales finies.

3.2.2.3. Conjonction de coordination (COO)

En vertu des propriétés caractéristiques données plus haut, *ainsi que, excepté, sauf* et *hormis* sont classés COO par leur faculté de constituer des deuxièmes termes de coordinations, même s'ils ont des emplois antéposés voisins de ceux des prépositions.

Le statut catégoriel de COO s'accompagne d'un traitement syntaxique particulier découlant de notre analyse de la coordination comme phénomène de listes syntaxiques (voir l'article de Kahane & *al*. dans ce volume). Les COO ont comme gouverneur la tête du deuxième conjoint et lui sont reliées par la relation MARK. Il s'agit là du seul cas d'une relation biunivoque entre catégorie et fonction.

Nous avons en effet choisi de classer *ainsi que, plutôt que* à la fois comme COO et comme CSU, alors que nous aurions pu considérer un seul classement comme CSU. Cette CSU fonctionnant comme subordonnant lorsqu'elle dépend d'un verbe par la fonction DEP et comme coordonnant lorsqu'elle occupe la fonction MARK dans une structure de coordination. Nous avons jugé peu intuitif pour l'utilisateur de présenter une CSU fonctionnant comme une COO : le principal critère distinguant COO et CSU étant la reprise possible par *que,* l'impossibilité d'avoir ***j'ai parlé à Marie ainsi qu'à Paul et qu'à Pierre* rendait difficile de maintenir le statut de CSU dans ces contextes de coordination. Cette solution peut être discutée.

3.2.2.4. Adverbe (ADV)

Pour respecter les habitudes de la majorité des utilisateurs, nous avons accepté l'héritage d'une catégorie adverbe présentant une grande hétérogénéité. Nous avons par exemple renoncé à en exclure les pro-groupes prépositionnels, ou, selon la formulation de M. Gross, les « pro-adverbiaux » *ainsi, là, ici…* Et nous y avons inclus des syntagmes prépositionnels figés comme *de fait, en revanche…* lorsque leur traitement comme item unique n'entraînait pas d'ambiguïtés avec des séquences régulières.

La classe rassemble des éléments morphologiquement et distributionnellement hétérogènes comme dans le classement traditionnel du Lefff :

- des items définis morphologiquement : adverbes terminés par *-ment* ;
- des items définis par une distribution particulière, entre auxiliaire et verbe sans rupture prosodique :
  - aspectuels et modaux comme *il a (encore / souvent / toujours / bien / peut-être) parlé à Marie*,
  - quantificurs comme *il a (trop / beaucoup / rien) mangé*,
  - négations comme *pas* et *jamais* ;
- des proformes substituts de SP qui, à la différence des items précédents, ont une distribution de SP : *ainsi, ailleurs, aujourd'hui, çà et là, comme ça, dehors, ensemble, ibidem, ici, jamais, là, partout, quelque part, quelquefois* ;
- On peut ajouter des PRO VRB comme *oui* (*il dit que oui*) ;
- des adverbes « connecteurs » : *donc, ensuite, par conséquent* ;
- des SP figés : on cherche à en limiter le nombre, mais on traite comme adverbes les groupes prépositionnels figés : *de fait, en fait, en somme*, *comme ça…* (Une cinquantaine d'items dont on trouvera la liste dans le lexique sur le site).

Nous avons également procédé à quelques analyses particulières. Par exemple, *jusque-là*, *jusqu'alors et jusqu'ici* sont classés comme adverbes composés dans le Lefff, tout comme *jusque* tout seul malgré l'absence d'emploi isolé (**j'attendrai jusque*). En nous fondant sur l'existence de séquences régulières PRE + ADV (*de là, par là*) et PRE + PRE (*près de là, par-dessus le mur*), nous avons décidé de classer *jusque* comme PRE. On a donc *jusqu'à* qui est analysé comme PRE + PRE et *jusqu'ici* comme PRE + ADV. Ce classement permet en outre de prendre en compte les séquences non standard attestées : *jusque la prochaine fois, jusque le bout de la rue*.

Par ailleurs, on accepte que plus d'adverbes que d'ordinaire aient une valence active. Outre les adverbes tirés d'adjectifs : *parallèlement à, conformément* à, on admet *loin de, près d*e, *autour de, tard dans la nuit.* Ainsi que les adverbes composés comme *à l'écart de*. La différence majeure avec les prépositions est que les adverbes, comme les adjectifs, n'admettent pas de « complément direct ».

3.2.2.5. Interjection (INT)

Contrairement à l'adverbe, les items classés ici :

- ne peuvent jamais être régis ;
- forment un tour de parole à interprétation autonome (peut initier un discours).

INT regroupe de fait les interjections traditionnelles et les particules discursives hors catégories : *euh, bon, ben etc., ouais*. Les incises comme *je crois, je vois, tu sais* gardent leurs catégories d'origine (CLS + VRB). Leur spécificité est traitée en syntaxe par l'intermédiaire de la relation DM (Voir l'article de Kahane & *al*. Dans

ce volume). C'est aussi le cas de *pardon, merde, putain* qui restent des noms dans leur emploi interjectif.

C'est en fait l'existence de formes que l'on ne peut assigner à une catégorie existante (*eh ! oh ! ouf, euh*) qui justifie la création de la catégorie INT. Sinon, on pourrait généraliser une analyse par relation syntaxique appliquée aux incises : on garde l'étiquette attestée par ailleurs : *merde* (N), *dommage* (N), *juste* (ADJ) *je sais* (VRB) et on les affecte soit d'une relation ROOT quand elles constituent un tour autonome, soit d'une relation DM quand elles sont insérées dans un tour de parole

3.3. Traitements particuliers d'un domaine

3.3.1. Les « connecteurs » comparatifs et consécutifs

Autant que, plus que, moins que, ainsi que, etc. posent des problèmes particuliers. On peut hésiter entre catégories PRE, COO, CSU. On a décidé de limiter le choix à COO et CSU et on a distingué trois emplois :

- les emplois où la séquence est ou pourrait être « décomposée » ;
- les séquences indécomposables homonymes des précédents ;
- les autres séquences indécomposables comme *ainsi que* et *plutôt que*.

Dans la première catégorie, on trouve des unités telles que *autant que, plus que, moins que, d'autant plus que* :

15. Il a travaillé autant qu'il s'est amusé.
16. Il a autant travaillé qu'il s'est amusé / que moi.
17. Il a d'autant plus travaillé qu'on l'a bien payé.

Dans ce cas, on décompose en ADV + CSU. Donc *que* est CSU dépendante du verbe ainsi que l'adverbe de quantité.

Dans la seconde catégorie, on trouve :

18. autant qu'on le fasse aujourd'hui
19. *autant peut-être qu'on le fasse aujourd'hui
20. tant qu'il restera /* tant alors qu'il restera

Dans ce cas, *autant que* et *tant que* sont CSU complexes.

Dans la troisième catégorie, on distingue deux contextes. 1) Dans le cas où la séquence gouverne un VRB, on aligne sur les CSU composées de type *alors que* :

21. Ainsi que je te l'avais dit, il est venu.
22. Il s'est comporté ainsi qu'on lui avait demandé.
23. Ainsi que les Grecs faisaient de la philo (de même) les Arabes faisaient des maths.
24. Plutôt qu'il aille à Paris, je préfère qu'il reste ici.

Dans *plutôt que d'aller à Paris je préfère rester*, on a traité l'ensemble *plutôt que de* comme une PRE. Mais dans *plutôt mourir*, *plutôt* est ADV comme dans *il est plutôt grand*. 2) Dans le cas où la locution gouverne un complément non phrastique équivalent d'une coordonnée :

25. Paul ainsi que Pierre viendront demain.
26. Je prendrai un livre plutôt qu'un cahier.

on a choisi de catégoriser l'ensemble COO, même dans le cas d'antéposition :

27. Ainsi que Pierre, Jean fait du surf.
28. Plutôt qu'un cahier, je prendrai un livre.

3.3.2. Les « exceptifs »

Les items ayant le même fonctionnement que *sauf, excepté, hormis* sont classés en tant que COO sur la base d'exemples comme :

29. Nous proposerons le poste à tous les bacheliers sauf à ceux qui n'ont pas d'expérience de l'animation.

L'analyse est étendue au cas de :

30. Sauf en cas d'urgence, il ne faut pas utiliser ce téléphone.

4. Traitement des mots composes

Les « mots composés » (selon la dénomination traditionnelle) ou les expressions « multi-mots » posent de nombreux problèmes de définition dont la solution ne fait pas l'unanimité parmi les linguistes. Cette question ne sera abordée ici que d'un point de vue pratique : choisir la solution qui minimise les erreurs d'analyse de l'analyseur.

4.1. Réduire les erreurs d'analyse dues au figement du composé

Comme nous l'avons annoncé dans l'introduction, le problème majeur que pose à l'analyseur la décision de figer en un seul « token » une suite d'items qui fonctionnent par ailleurs comme des tokens séparés est que la séquence libre ne sera pas reconnue par l'analyseur. Ainsi si l'on décide de figer la suite *bien que* en un seul mot analysé comme conjonction de subordination dans le lexique, l'analyseur ne pourra pas donner une analyse correcte de la suite *bien que* dans la séquence : *je sais bien que ce n'est pas juste.*

Pour des raisons de temps, nous avons choisi de traiter le problème des composés « fonctionnels » appartenant aux catégories « fermées » de PRE, CSU, PRO, PRQ. Nous considérons en outre que les différences de tokénisation dans le domaine des catégories « ouvertes » (NOM, ADJ, VRB…) a moins d'incidence sur l'analyse en dépendances. Pour ces dernières catégories, nous avons suivi le Lefff.

4.2. Traitement des composés fonctionnels

Notre attitude pragmatique consiste à croiser deux problèmes concernant les composés : d'une part, définir les critères linguistiques selon lesquels une séquence de plusieurs tokens est considérée comme un seul mot dans le lexique, et d'autre part, adapter ces critères en fonction des conditions de fonctionnement de l'analyseur. Nous traitons les composés comme des séquences possédant un gouverneur qui est le seul à avoir un lien de dépendance avec le contexte. Les autres éléments de la séquence sont reliés au gouverneur par un lien spécifique MORPH représentant l'idée que ces éléments sont des extensions du morphème tête. Nous précisons ci-dessous les critères qui sont utilisés pour établir un lien MORPH.

4.2.1. Critères linguistiques

L'intuition de non compositionnalité sémantique de la séquence n'est pas un critère suffisant pour la déclarer composée. Elle peut fonctionner comme une heuristique, mais doit être accompagnée de critères formels. Nous prendrons l'exemple des suites comportant une CSU.

La séquence multi-mots peut être considérée ou pas comme une suite syntaxique régulière dans notre modèle. Par exemple, *pour que* ne sera pas composé parce que

nous analysons la séquence comme la suite régulière PRE + CSU, la CSU étant un dépendant régulier de PRE. Idem pour *dans le but que*. Au contraire*, afin que*, *mis à part que* ne sont pas des suites régulières, car on ne peut attribuer une catégorie à *afin,* et *à part + que* ne sont pas des dépendants possibles pour *mis.* On a donc un seul mot dans le lexique. Nous considérons de la même façon les séquences suivantes comme non régulières : *à condition que, à force que*, *toujours est-il que.*

Pour ce qui est de *bien que*, on constate qu'il peut être analysé sur le modèle de *heureusement que*, mais sa distribution est celle d'une conjonction de subordination et non celle d'un adverbe (*bien* ne peut être root). Nous en faisons un candidat à la composition.

Les séquences candidates à la composition sont analysées au moyen du lien MORPH qui va d'un item choisi comme tête aux autres items du composé. Notre originalité tient à la volonté de motiver le plus possible le choix de la tête. Nous distinguons les composés endogènes où on peut considérer que l'ensemble a la distribution d'un des composants (*bien que* composé a la distribution d'une conjonction de subordination comme *que*) et les composés exogènes où l'ensemble a une distribution différente de ses parties (ex : *c'est pourquoi*). Nous avons constaté empiriquement que pour les CSU et les PRE, il y avait très peu d'exogènes. La situation est différente pour les COO. *Ainsi que, plutôt que, au même titre que, c'est-à-dire,* qui sont classés COO, sont des composés exogènes dans ce fonctionnement.

4.2.2. Critères liés au fonctionnement de l'analyseur

Notre intention initiale était d'annoter selon ces principes le corpus d'entraînement de l'analyseur et de lui laisser le soin de définir lui-même en fonction de son entraînement l'attribution du lien MORPH et donc des composés du corpus à annoter. Cette technique devait en particulier permettre à l'analyseur de choisir entre l'analyse en composé et l'analyse standard dans le cas des séquences potentiellement ambigües : *bien que, alors que, pour autant que…*

Nous avons alors rencontré le problème suivant : la faible fréquence de certaines séquences pouvait rendre difficile l'apprentissage des différences d'analyse. Nous avons observé de fait que certaines séquences ont un fonctionnement préférentiel : ainsi *tant que* est majoritairement décomposé (80% des occurrences sur notre échantillon), tandis que *alors que* est majoritairement un composé (88% des occurrences). Nous avons donc utilisé cette constatation pour simplifier l'analyse. Nous avons intégré directement dans le lexique des CSU les cas où l'ambiguïté était prévisiblement faible *: une fois que, à condition que, à part que.*

5. Evaluation et perspectives

Précisons tout d'abord que les propositions d'améliorations que nous allons faire seront évaluées et intégrées aux ressources présentes sur la plateforme Orféo par le suivi qu'assurera l'équipe TALEP (Traitement automatique de la langue écrite et parlée).

Les mesures de performances obtenues à partir de ce jeu d'annotation présentent une précision de 97,19 %. Ce résultat est satisfaisant, mais il peut sans doute être amélioré. Si nous nous plaçons dans une perspective de construction de requêtes, il n'a pas en soi de valeur absolue : il doit être rapporté aux résultats de l'analyse

automatique en dépendances. Ce sont ces résultats qu'il faut prioritairement améliorer et toute modification du jeu d'étiquettes et de la composition du lexique n'est intéressante que dans la mesure où elle ne les dégrade pas. Quelles sont maintenant les pistes d'amélioration que nous envisageons ? Nous avons vu que pour certaines des parties du discours les choix ont été faits en sacrifiant la cohérence à la facilité d'utilisation. Trois chantiers pourraient être ouverts.

1. Dans la perspective de limiter les doubles classements, exploiter jusqu'au bout la possibilité de dissocier la catégorie et la fonction, notamment dans le traitement des items comme *ainsi que, plutôt que*.

2. Plus généralement, améliorer le traitement des expressions multi-mots. Cette piste pourra bénéficier des apports du projet PARSEME dans lequel est engagé notre groupe de travail. L'idée est de développer le plus possible la notion de *suite régulière*, notamment en établissant une complémentarité entre une approche syntaxique et une approche sémantique de la notion de figement.

3. Enfin, un défi important serait de vérifier si des avancées récentes dans la théorie des parties de discours peuvent être utilisées pour modifier le tagset. Les discussions précédentes nous amènent avant tout à remettre en cause la catégorie de l'adverbe dont nous avons assumé, pour des raisons pratiques, la composition *ad hoc*. On pourrait imaginer deux étapes. L'une dans la quelle on réduit la catégorie à un noyau dur de lexèmes à distribution homogène. L'autre dans laquelle on élimine purement et simplement cette catégorie.

Pour le premier objectif, il serait immédiatement possible de reverser dans d'autres catégories certaines sous classes des adverbes actuels qui y trouveraient mieux leur place.

- Les adverbes en *-ment* peuvent être considérés comme des formes flexionnelles d'adjectifs avec lesquels ils partagent beaucoup de propriété valencielles.
- Les adverbes qui ont des propriétés de proformes substituts de SP (*là, ainsi, aussi…*) peuvent être réassignés à une catégorie proforme qui serait une extension de la catégorie PRO comparable à ce qui a été fait pour CLI et PRQ.
- Les adverbes quantifieurs ou de degrés constituent des catégories fermées. Il serait possible de les rapprocher d'autres catégories fermées comme celle des DET : ils pourraient être classés comme des déterminants d'adjectifs (***très*** *beau*) ou de prépositions (***juste*** *à côté de ça*).

Il resterait pour constituer la catégorie adverbe, les adverbes « connecteurs » comme *donc, enfin…* et les adverbes exprimant le temps, l'aspect et les modalités (*bien, encore, souvent*). Cette catégorie serait alors caractérisée par une propriété topologique signalée plus haut : apparaître entre l'auxiliaire et le verbe sans rupture intonative (Bonami & Godard 2007).

Si ce critère apparaît trop fragile pour définir une catégorie, il reste encore la possibilité d'éliminer totalement la catégorie adverbe. On pourrait le faire en suivant J. Emonds (2000, chap. 3) qui propose de généraliser le critère de la prémodification pour définir seulement 4 parties du discours majeures (N, V, A, Pré), dans le cadre d'une première distinction entre catégories lexicales ouvertes et catégories fonctionnelles fermées.

Bien évidemment, l'adoption de tout nouveau système serait liée à l'amélioration produite dans les performances de l'analyseur. De façon intéressante, l'efficacité du programme informatique pourrait constituer un critère de validation des systèmes de

parties du discours concurrents, ce qui établirait des liens nouveaux entre analyse linguistique et construction d'outils de traitement automatique.

Annexe : liste des étiquettes avec leur nom complet

ADJ (adjectifs qualificatifs) : *méchant*, *petit*, *long*, etc.
ADN (adverbes de négation) : *pas*, *jamais*, *nullement*, *guère*, *plus*, etc.
ADV (adverbes) : *savamment*, *peut-être*, *in extremis*, *très*, *environ*, etc.
CLI (autres clitiques) : *te*, *lui*, *-le*, *-y*, *en*, *-leur*, *nous*, etc.
CLN (clitique de négation) : *ne*
CLS (clitiques sujets) : *tu*, *elles*, *vous*, *-vous*, *c'*
COO (conjonctions de coordination) : *et, ou, alias*, *mais encore*, *voire*, *puis*, etc.
CSU (conjonctions de subordination) : *alors que*, *lorsque*, etc.
DET (déterminants) : *cette, certains, quelques*, *un*, etc.
INT (interjections) : *hein, ben, allô*, *pfff*, *no comment*, *parbleu*, etc.
NOM (noms) : *diplodocus*, *topinambour*, *Google*, etc.
NUM (nombres) : *six*, *treize*, *quatorze, vingt-cinq,* etc. (mais pas *soixante et un*)
PRE (prépositions) : *de, des, parmi*, *pour cause de*, *par delà*, *outre*, etc.
PRO (pronoms) : *moi, celles*, *les tiens*, *plusieurs*, *nul*, *pas grand-chose*, etc.
PRQ (pronoms interrogatifs-relatifs) : *combien est-ce que*, , *pourquoi*, *que*, etc.
VNF (verbes à l'infinitif) : *tenir*, *poindre*, *jouer*, *entendre*, etc.
VPP (verbes au participe passé) : *tenu*, *point*, *joué*, *entendu*, etc.
VPR (verbes au participe présent) : *tenant*, *poignant*, *jouant*, *entendant*, etc.
VRB (verbes à la forme finie) : *tiens, poignent, joueraient, entendissions*, etc.

Références

AARTS B. (2004), « Modelling linguistic gradience », *Studies in Language* 28, 1-49.

BONAMI O. & GODARD D. (2007), « Quelle syntaxe, incidemment, pour les adverbes incidents ? », *Bulletin de la Société de Linguistique de Paris 102*, 255-284.

CROFT W. (2000), « Parts of speech as typological universals and as language particular categories », in P. M. Vogel & B. Comrie (eds), *Approaches to the typology of word classes,* Berlin : Mouton de Gruyter, 65-102.

EMONDS J. (2000), *Lexicon and grammar: the english syntacticon*, Studies in generative grammar 50, Berlin : Mouton de Gruyter.

JESPERSEN O. (1924), *The philosophy of grammar*, London : George Allen and Unwin.

NASR A., RAMISH C., DEULOFEU J. & VALLI A. (2015), « Joint dependency parsing and multiword expression tokenization », *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics* Vol. 1, Beijing China, Association for Computational Linguistics, 1116-1126.

SAGOT B., CLÉMENT L., DE LA CLERGERIE É. & BOULLIER P. (2006), « The Lefff 2 syntactic lexicon for French: architecture, acquisition, use », *Proceedings of the 5th Language Resources and Evaluation Conference (LREC'06)*, Genova, Italie.

SAGOT B. et DANLOS L. (2007), « Améliorer un lexique syntaxique à l'aide des tables du lexique-grammaire – Constructions impersonnelles », *Cahiers du Cental.*

SAGOT B. et DANLOS L. (2008), « Méthodologie lexicographique de constitution d'un lexique syntaxique de référence pour le français », *Actes du colloque Lexicographie et informatique : bilan et perspectives*, Nancy, France.

Ressource collective : Recensement des lexiques de parties de discours utilisés pour le français http://french-postaggers.tiddlyspot.com/