Construction automatique d'un large corpus libre annoté morpho-syntaxiquement en français

Nicolas Hernandez, Florian Boudin

Construction d'un large corpus écrit libre annoté morpho-syntaxiquement en français

Nicolas Hernandez Florian Boudin
Université de Nantes
nicolas.hernandez@univ-nantes.fr, florian.boudin@univ-nantes.fr

RÉSUMÉ

Cet article étudie la possibilité de créer un nouveau corpus écrit en français annoté morpho-syntaxiquement à partir d'un corpus annoté existant. Nos objectifs sont de se libérer de la licence d'exploitation contraignante du corpus d'origine et d'obtenir une modernisation perpétuelle des textes. Nous montrons qu'un corpus pré-annoté automatiquement peut permettre d'entraîner un étiqueteur produisant des performances état-de-l'art, si ce corpus est suffisamment grand.

ABSTRACT

Construction of a Free Large Part-of-Speech Annotated Corpus in French

This paper studies the possibility of creating a new part-of-speech annotated corpus in French from an existing one. The objectives are to be free from the restrictive licence of the source corpus and to obtain a perpetual modernisation of texts. Results show that it is possible to train a state-of-the-art POS-tagger from an automatically tagged corpus if this one is large enough.

MOTS-CLÉS : corpus arboré, construction de corpus, étiquetage morpho-syntaxique.

KEYWORDS: French treebank, Building a corpus, Part-of-Speech Tagging.

1 Introduction

L'entraînement et le test de systèmes statistiques de Traitement Automatique des Langues (TAL) requièrent la disponibilité de larges corpus annotés (Hajičová et al., 2010). Force est de constater que la communauté scientifique est pauvre en corpus écrits en français librement accessibles, annotés en quantité et en qualité suffisantes avec des analyses linguistiques structurelles (segmentation des textes en titres, paragraphes, phrases et mots), morpho-syntaxiques (parties du discours, lemme, genre, nombre, temps...) et syntaxiques (en constituants et en dépendances) qui constituent les pré-traitements de la plupart des applications du TAL. Nous reprenons ainsi à notre compte des propos énoncés près de dix ans plus tôt dans (Salmon-Alt et al., 2004). Dans cet article nous nous intéressons à l'entraînement d'étiqueteurs morpho-syntaxiques pour traiter des écrits en français ainsi qu'à la construction des corpus annotés associés.

Parmi les corpus écrits annotés et en français que nous recensons, nous comptons *PAROLE* [1] et *MULTEXT JOC* [2] (Véronis et Khouri, 1995), le *French Treebank* (P7T) (Abeillé et al., 2003), la base

1. http://catalog.elra.info/product_info.php?products_id=565
2. http://catalog.elra.info/product_info.php?products_id=534

FREEBANK (Salmon-Alt et al., 2004) et le récent corpus *Sequoia* [3] (Candito et Seddah, 2012). Excepté la *FREEBANK*, ces corpus sont toujours accessibles aujourd'hui via un guichet sur le Web. La *FREEBANK*, dont la motivation était le recueil collaboratif, la construction et le partage de corpus libres annotés en français a malheureusement disparu dans les limbes du Web [4] du fait de la difficulté d'acquisition de textes libres et du coût de réalisation d'une telle entreprise.

Le P7T [5] est probablement le corpus annoté le plus utilisé et le plus référencé, et ce essentiellement pour trois raisons : il est libre d'usage pour des activités de recherche, il bénéficie d'une analyse multi-niveaux (de la structure textuelle à la structure syntaxique en passant par des annotations en morphologie) et il compte près du double de mots annotés que tous les autres corpus disponibles réunis. En pratique ce corpus se compose d'articles journalistiques issus du journal Le Monde écrits dans les années 90, soit plus de 500 000 mots annotés. Ainsi (Candito et al., 2010a) utilisent la structure en constituants du P7T pour construire une structure en dépendances et permettre l'entraînement d'analyseurs syntaxiques statistiques en dépendance du français (Candito et al., 2010b). (Sagot et al., 2012) l'enrichissent avec des annotations référentielles en entités nommées. Tandis que (Danlos et al., 2012) projettent de l'utiliser comme base d'annotations discursives.

Il y a néanmoins quelques problèmes associés à l'utilisation du corpus P7T dans une optique de développement de systèmes statistiques de TAL.

1. Le premier problème concerne la faible adéquation du modèle théorique linguistique avec la tâche d'entraînement à laquelle on le destine. (Schluter et van Genabith, 2007; Crabbé et Candito, 2008) montrent qu'en remaniant certaines annotations syntaxiques et le jeu d'étiquettes, il est possible d'améliorer les performances des systèmes entraînés avec ce corpus. Un autre aspect du problème porte sur la notion de mots composés définie par les auteurs. Celle-ci est très large et a pour conséquence de rendre difficilement reproductible la segmentation du P7T par un système automatique non entraîné sur cette ressource. Cette conséquence conduit à s'interroger sur la pertinence d'utiliser des modélisations construites sur ce corpus pour traiter d'autres corpus. (Candito et Seddah, 2012), par exemple, décident de restreindre cette définition et d'aborder le traitement des formes les plus ouvertes (composés nominaux et verbaux) qu'au niveau syntaxique. En comparaison, le *Penn Treebank* [6], qui constitue la référence pour l'anglais-américain (Marcus et al., 1993; Gabbard et al., 2006), favorise un découpage en mots simples [7] en privilégiant la rupture pour les mots joints. Il est néanmoins important de rappeler que le P7T a initialement été créé avec une motivation différente de la nôtre aujourd'hui à savoir la construction de ressources lexicales de type dictionnaire [8].

2. Le second problème est plus technique et concerne la relative inadéquation du schéma XML de représentation des annotations pour des tâches automatiques ainsi que le manque de consistance de la structure d'annotation. La représentation des amalgames en deux éléments XML distincts qui se retrouvent distribués dans différentes configurations selon qu'ils se produisent en partie dans un mot composé est une situation difficile à traiter automatiquement car elle oblige à énumérer tous les cas possibles. Certains éléments n'ont pas systématiquement tous leurs attributs, d'autres ont des noms d'attribut erronés... Ces

3. https://www.rocq.inria.fr/alpage-wiki/tiki-index.php?page=CorpusSequoia
4. http://web.archive.org/web/20081215041844/http://freebank.loria.fr/
5. http://www.llf.cnrs.fr/Gens/Abeille/French-Treebank-fr.php
6. http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC99T42
7. http://www.cis.upenn.edu/~treebank/tokenization.html
8. http://www.llf.cnrs.fr/Gens/Abeille/guide-morpho-synt.02.pdf

inconsistances sont relevées dans de nombreux travaux (Arun et Keller, 2005; Schluter et van Genabith, 2007; Green et al., 2011; Candito et Seddah, 2012; Boudin et Hernandez, 2012) qui militent en faveur d'une amélioration voire d'une réorganisation de la structure du P7T avant de pouvoir l'utiliser dans toute étude sérieuse.

3. Généralement les mêmes auteurs ont aussi observés des inconsistances au niveau d'annotations. Certains, comme (Schluter et van Genabith, 2007; Boudin et Hernandez, 2012), mettent en oeuvre des techniques automatiques pour détecter et corriger des erreurs d'étiquetage morpho-syntaxique. Le nombre d'erreurs de ce type est souvent minime ramené au nombre de mots annotés. Sur les 628 767 tokens mots considérés dans (Boudin et Hernandez, 2012) par exemple, seulement 169 ont été considérés comme ayant une erreur d'étiquette. Le corpus étant construit semi-automatiquement (d'abord analysé automatiquement puis validé manuellement), ce type de problème illustre le fait que la validation humaine ne garantie pas l'absence d'erreurs sur un large corpus.

4. Le quatrième problème que nous relevons résulte d'un parti pris que nous prenons [9]. Nous estimons en effet que sa licence d'exploitation n'est pas adaptée pour favoriser son utilisation dans le monde de la recherche. Bien que la licence permette des utilisations avec des outils propriétaires et à des fins commerciales moyennant finance, elle n'autorise pas la modification et la diffusion libre des modifications du corpus. Cela a pour principale conséquence de ralentir voire de décourager les contributions extérieures et l'amélioration de la ressource (par exemple pour corriger les problèmes précédemment cités).

5. Les données annotées sont des textes mono-genres vieux de près de vingt ans encodés en ISO-8859-1. On peut se poser la question de la robustesse et de la précision des systèmes entraînés sur ceux-ci pour traiter des textes plus récents (qui présentent de nouveaux phénomènes linguistiques et des caractères encodés en UTF-8, qui est le standard de facto aujourd'hui pour encoder des textes en français) et/ou de genre différent.

6. Même s'il constitue le plus gros corpus annoté disponible pour le français, on peut s'interroger sur la représentativité d'un corpus d'un demi-million de mots pour la construction de systèmes automatiques. A titre de comparaison, le *Penn Treebank* compte en corpus écrits près de 2,4 millions de mots annotés morphologiquement et syntaxiquement et couvrent le domaine journalistique (*Wall Street Journal*) et l'anglais général (*Brown*).

Comparativement, *PAROLE* et *MULTEXT JOC* ont aussi des licences restrictives, le *Sequoia* offre quant à lui le plus de libertés [10] aux utilisateurs. Aucun des corpus n'est de taille comparable à celle du P7T. Ils comptent respectivement 250 000, 200 000 et 72 311 mots annotés morpho-syntaxiquement. Excepté en partie pour le *Sequoia*, les textes datent des années 80 et 90.

Dans cet article, nous ré-ouvrons la question de la construction de corpus annotés libres en français. Une conjoncture à la fois sociétale, politique, technique et scientifique nous y conduit. En effet nous bénéficions aujourd'hui d'au moins deux sources de contenu libres et multilingues, en croissance perpétuelle et comptant déjà plusieurs millions de mots, à savoir les projets de

9. Nous nous situons dans une démarche de recherche scientifique «ouverte» (Nielsen, 2011).

10. LGPL-LR (Lesser General Public License For Linguistic Resources). Les auteurs ne précisent pas l'objet désigné par la licence. Celle-ci doit se restreindre aux annotations produites et ne peut comprendre les textes. Le corpus est composé de textes de quatre origines. On note que le journal Est Républicain diffusé par le CNRTL est sous licence CC-BY-NC-SA 2.0 FR qui par sa clause de non-diffusion commerciale s'oppose à la LGPL-LR. La licence de wikipedia (CC-BY-SA 3.0) ne semble pas contredire cette licence. La licence de Europarl et de EMEA manque de précision sur les droits d'usage mais autorise la reproduction.

la Wikimedia Foundation [11] (*Wikipedia*, *Wikinews*...) et les actes du Parlement Européen [12] (*Europarl*) tels que remaniés par (Koehn, 2005). Nous nous intéresserons ici aux écrits en français de *Wikinews* et de *Europarl*. La version en français de Janvier 2013 de *Wikinews* compte plus de 28 000 articles d'actualité (soit plus de 2,5 millions de mots sur près de 90 000 phrases) et couvre une période s'étalant de Janvier 2005 à nos jours. La section en français de la version 7 (mai 2012) du corpus *Europarl* compte, quant à elle, plus de 61,5 millions de mots (plus de 2 millions de phrases) et couvre une période s'étalant de 1996 à 2011. Les textes du premier sont disponibles sous licence [13] *Creative Commons Attribution 2.5 (CC-BY 2.5)* (les versions antérieures à Septembre 2005 sont dans le domaine public) qui permet à l'utilisateur d'utiliser, de modifier et de diffuser la ressource et ses modifications comme il le souhaite moyennant l'obligation d'en citer l'auteur. Les textes du second sont libres de reproduction [14].

Dans les sections suivantes, nous nous interrogeons sur la possibilité d'exploiter des données pré-annotées automatiquement pour construire un système ayant des performances similaires à des systèmes entraînés sur des données validées manuellement. Nous proposons notamment d'observer comment la taille des données pré-annotées automatiquement peut jouer un rôle dans la performance d'un étiqueteur morpho-syntaxique entraîné sur celles-ci.

2 Cadre expérimental

Dans cette section, nous présentons les données, le jeu d'étiquettes et l'étiqueteur que nous utilisons (section 2.1). Nous présentons aussi les pré-traitements opérés sur les données pour les exploiter (sections 2.2 et 2.3) ainsi que le protocole d'évaluation de nos expériences (section 2.4).

2.1 Données, jeu d'étiquettes et étiqueteur

Pour nos expérimentations nous utilisons tour à tour le corpus P7T comme données d'entraînement et de test. Le corpus *Sequoia* est aussi utilisé selon les expériences.

Nous utilisons les parties en français du *Wikinews* et d'*Europarl* comme données non étiquetées. Nous filtrons les phrases courtes (i.e. inférieures à 5 tokens) de chaque document et nettoyons la syntaxe wiki de *Wikinews*. L'ensemble de données ainsi généré possède plusieurs avantages. Tout d'abord, *Wikinews* est du même genre que le P7T (journalistique). La différence de genre avec *Europarl* permet de discuter de la portabilité de l'approche à des genres différents. Ensuite ces corpus possèdent une taille bien supérieure au P7T ; environ quatre fois supérieure pour *Wikinews* et soixante fois pour *Europarl*. Enfin la licence associée à ces ressources permettent de les distribuer librement accompagnées des annotations que nous générons.

Le jeu de catégories morpho-syntaxiques que nous utilisons est celui mis ou point par (Crabbé et Candito, 2008), contenant 28 catégories qui combinent différentes valeurs de traits morpho-syntaxiques du P7T. Outre le fait que ce jeu soit plus complet que les catégories du P7T, qui

11. http://wikimediafoundation.org
12. http://www.statmt.org/europarl/
13. http://dumps.wikimedia.org/legal.html
14. «Except where otherwise indicated, reproduction is authorised, provided that the source is acknowledged.» http://www.europarl.europa.eu/guide/publisher/default_en.htm

elles sont au nombre de 13, les auteurs montrent que les performances d'un étiqueteur entraîné sur de telles annotations sont meilleures. Par ailleurs, son utilisation facilite l'accès à d'autres ressources tels que les analyseurs syntaxiques statistiques en dépendance du français qui ont déjà été développés à partir de ce jeu d'étiquettes [15] (MaltParser, MSTParser, Berkeley Parser) (Candito et al., 2010b). Par la suite nous ferons référence à ce jeu d'étiquette par le nom P7T+. Par abus ce nom désignera aussi le corpus P7T avec des étiquettes converties en P7T+.

En ce qui concerne l'étiqueteur morpho-syntaxique que nous avons utilisé pour nos expériences, il s'agit de la version 3.1.3 du *Stanford POS Tagger* (Toutanova et al., 2003). Ce système utilise un modèle par maximum d'entropie, et peut atteindre des performances au niveau de l'état-de-l'art en français (Boudin et Hernandez, 2012). Nous utilisons un ensemble standard [16] de traits bidirectionnels sur les mots et les étiquettes.

2.2 Segmentation en mots

Le P7T fournit des analyses linguistiques qui reposent sur une segmentation en mots simples et en mots composés. Les mots composant les composés (nous appelons «mots composants» les mots qui composent les mots composés) sont signalés mais seulement un sous-ensemble bénéficie d'une catégorie grammaticale et aucun d'eux ne bénéficie des autres traits (sous-catégorie, flexions morphologiques et lemme). Excepté le lemme, ces traits sont requis pour la conversion en P7T+.

La notion de composé dans le P7T est très large (cf. note 8). La composition se justifie par des critères aussi bien graphiques que morphologiques, syntaxiques et sémantiques. La segmentation en unités lexicales n'est pas un problème trivial. De nombreuses marques de ponctuation (apostrophe, virgule, tiret, point et espace) sont ambiguës, et suivant la situation, jouent le rôle de joint ou de séparateur. Cela conduit la majorité des systèmes de segmentation (Benoît et Boullier, 2008; Nasr et al., 2010; Constant et al., 2011) à exploiter, en complément de règles générales, des listes de formes finies ou régulières à considérer comme unités lexicales. La segmentation en composés du P7T résulte d'un processus d'annotation à la fois manuel et à base de lexiques non précisément référencés. Outre la difficulté à reproduire automatiquement cette segmentation, il n'y a pas d'enjeu à chercher à le faire car celle-ci est avant tout ad hoc à une période et un genre de textes. Motivés par la volonté d'entraîner des analyseurs robustes afin de pouvoir traiter des textes pour lesquels des dictionnaires de mots composés ne seraient pas disponibles, nous avons souhaité nous abstraire au maximum de la notion de composé du P7T. Nous n'avons ainsi considéré comme unités lexicales que les composés consistant en des unités graphiques exemptes d'espace ou ceux consistant en des formes numérales régulières (e.g. «20 000», «50,12», «deux cent vingt-et-un»), lesquelles peuvent admettre des espaces. Certains mots composants sont donc amenés à être considérés comme unités lexicales. Il en découle le besoin de déterminer les traits morpho-syntaxiques manquants de ceux-ci afin de pouvoir leur affecter une étiquette P7T+ (cf. section 2.3). Le P7T compte 6 791 lemmes distincts de mots composés qui ne sont pas des unités graphiques (i.e. ne contenant pas d'espace) soient 26 648 occurrences. 1 892 de ces lemmes de mots composés ont au moins un de leur composant sans étiquette grammaticale. Cela représente 7 795 occurrences. 1 106 n'ont aucune étiquette à leurs composants.

Pour les données autres que le P7T, nous utilisons dans nos expériences le segmenteur KEA [17].

15. `http://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html`
16. Nous avons utilisé la macro `generic,naacl2003unknowns` décrite dans (Toutanova et al., 2003).
17. `https://github.com/boudinfl/kea`

2.3 Révision et extension du P7T

Afin de faciliter le traitement automatique ultérieur du P7T, nous réalisons des opérations de révision et d'extension de la forme et du contenu. Nous envisageons à terme la construction de modélisations de toutes les informations disponibles dans le P7T pour des systèmes prédictifs statistiques. Nous avons jugé compliquées les situations où l'extraction de certaines informations nécessite des travaux d'analyse dédiées (que cela soit des analyses de valeurs d'attributs ou des manipulations de la modélisation objet des documents (DOM) pour obtenir différents fragments d'une même information).

Concernant les opérations visant la validation [18], l'homogénéisation et la simplification de la structure XML des documents (second type de problèmes recensé à la section 1), nous avons par exemple fusionné les éléments XML composant les amalgames [19] en un seul élément de manière similaire à (Candito et Seddah, 2012) (19 489 fusions). Nous avons explicité les caractéristiques morphologiques de chaque mot par des attributs propres (genre, nombre, temps, personne...). Nous avons fait diverses corrections pour valider les documents comme l'ajout d'attributs manquant (e.g. 3166 attribut `compound` ajoutés) et le renommage d'attribut (e.g. 3 726 attributs `cat` corrigés en `catint`).

Concernant les opérations de modification de contenu, la segmentation en tokens mots originale et le contenu textuel du P7T ont été épargnés. De même, les corrections d'erreurs triviales d'étiquetage ont été considérées à la marge pour cette étude. Les opérations se sont concentrées d'une part sur la détermination des traits morpho-syntaxiques (catégorie grammaticale, sous-catégorie, flexion morphologique et lemme) des mots composant les mots composés, et d'autre part sur l'attribution à chaque mot de l'étiquette grammaticale du jeux d'étiquettes du P7T+ correspondant à ses attributs. Ces deux types d'opérations, dont le détail est présenté respectivement dans les deux paragraphes suivants, visent à traiter le premier type de problèmes recensé à la section 1 ; en particulier la détermination des traits morpho-syntaxiques est une étape nécessaire à l'affectation d'une étiquette P7T+ aux mots composants (cf. section 2.2).

Le processus de détermination des traits manquants pour les mots composants repose sur l'observation des séquences de traits associées aux occurrences des composés, aux séquences de mots simples correspondant aux composants des composés, ainsi que sur l'observation des traits associés individuellement à chaque mot du corpus. Notre approche tente d'abord une résolution avec des statistiques globales et s'appuie ensuite sur des traits locaux au composé en cas d'ambiguïté au niveau global. Sur les 1 892 lemmes de composés incomplets que nous observons, nous proposons une solution à 1 736 (3 009 occurrences).

Le processus d'attribution à chaque mot d'une étiquette du P7T+ exploite les traits catégorie, sous-catégorie et flexion morphologique des mots. Pour ce faire, nous nous sommes appuyés sur la table de conversion énoncée par (Crabbé et Candito, 2008) ainsi que sur la documentation de l'étiqueteur morpho-syntaxique MELT (Denis et Sagot, 2010) pour compléter quelques règles manquantes [20]. 31 règles réalisent la conversion. Sur les 679 584 mots (simples, composés et composants) que compte le P7T, la procédure attribue une étiquette P7T+ à 664 240 mots ;

18. Seulement 27 des 44 fichiers composant la section *tagged* (étiqueté grammaticalement) de la version de Janvier 2012 sont valides (c'est-à-dire vérifient la spécification définie par le schéma NG fourni par les auteurs.)

19. Les amalgames sont des unités lexicales décrite par une unité graphique mais composés deux catégories grammaticales (e.g. "du" pour "de+le", "auxquel" pour "à+lequel").

20. Un mot de catégorie "Nom" et de sous-catégorie "cardinal" (million, huit, 2001...) est converti en nom commun. L'étiquette "préfix" ne change pas, comme celle des amalgames après fusion de ses sous-éléments.

15 344 sont donc indéfinis. Nos règles de conversion, testées sur les annotations P7T du corpus *Sequoia*, produisent les mêmes [21] annotations P7T+ que le corpus met aussi à disposition.

En pratique les différentes opérations ont été mises en oeuvre via des règles [22] plus ou moins générales exprimées sur le DOM des documents. Les opérations de comptage requises par certaines stratégies ont été réalisés sur tout le corpus et non seulement sur chaque document.

2.4 Protocole d'évaluation

Notre objectif est d'évaluer les performances d'un étiqueteur morpho-syntaxique construit sur des données pré-annotées automatiquement par rapport à un étiqueteur construit sur des données validées manuellement. Notre méthodologie est présentée à la figure 1. La première étape consiste à produire l'ensemble de données d'entraînement. Pour cela, nous utilisons le *Stanford POS tagger* avec un modèle entraîné sur le P7T+ pour annoter un large corpus de données non-étiquetées. Cet ensemble de données est noté CORPUS^{pos} après qu'il ait été étiqueté morpho-syntaxiquement. Nous l'utilisons alors dans une deuxième étape pour entraîner un nouveau modèle. La performance du modèle créé à partir de CORPUS^{pos} est ensuite évaluée sur le P7T+.

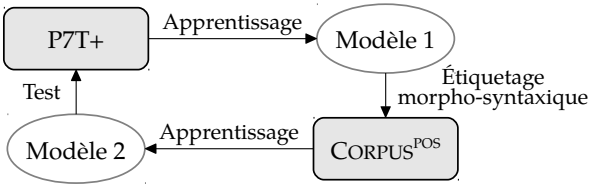

FIGURE 1 – Apprentissage d'un modèle à partir de données automatiquement annotées.

Afin d'étudier l'impact de la taille du corpus d'entraînement sur la performance de l'étiquetage morpho-syntaxique, nous avons entraîné différents modèles en utilisant des portions de CORPUS^{pos} représentant un facteur x du nombre de phrases de P7T+. Ici, nous utilisons *Wikinews* et *Europarl* en français comme CORPUS. Pour *Wikinews*, nous avons testé les facteurs allant de 1 à 4 fois le nombre de phrases de P7T+ (4 étant la limite que nous pouvions atteindre avec le nombre de phrases contenu dans *Wikinews*). Pour *Europarl*, nous avons exploré jusqu'au facteur 16.

Trois mesures d'évaluation sont considérées comme pertinentes pour nos expériences : la précision sur les tokens, la précision sur les phrases (nombre de phrases dans lesquelles tous les tokens ont été correctement étiquetés par rapport au nombre de phrases total) et la précision sur les mots inconnus (calculée à partir des tokens n'apparaissant pas dans l'ensemble d'entraînement).

21. 108 mots obtiennent une étiquette différente de celle attribuée par les auteurs du *Sequoia*, à savoir une étiquette désignant une valeur indéfinie. En y regardant d'un peu plus près nous avons constaté que cela concernait en fait 22 formes distinctes et que ces formes étaient ambiguës et pouvaient correspondre à des noms communs ou bien à des adverbes négatifs (e.g. 34 «personnes ?», 37 «points ?»). En creusant davantage, nous avons constaté un problème d'annotation. Ces mots étaient annotés en tant que nom (catégorie «N») mais possédaient une sous-catégorie «NEG» propre aux adverbes. La description incomplète de certains traits semblent être aussi la raison de l'attribution d'une étiquette indéfinie. C'est le cas de verbes («aboutisse, «agrandisse», «remplisse») dont le mode n'est pas précisé. Indirectement notre système a permis ainsi de détecter des erreurs d'inconsistances dans le *Sequoia*.

22. L'outil de révision et d'extension est librement disponible sur `https://sites.google.com/site/nicolashernandez/resources`

3 Expériences

Cette section présente les expériences que nous avons menées. Nous rapportons d'abord la performance d'un étiqueteur état-de-l'art construit sur des données manuellement validées (section 3.1). Puis nous rapportons les performances observées pour différentes tailles de données d'entraînement annotées automatiquement et ce pour des corpus d'entraînement de deux genres différents (sections 3.2 et 3.2). Enfin nous rapportons les performances de ces étiqueteurs construits sur des données non validées sur un corpus sans aucun lien de filiation connu (section 3.3). Le modèle et les traits d'entraînement de ces étiqueteurs sont présentés à la section 2.1.

3.1 Performance d'un étiqueteur état-de-l'art

La première expérience que nous avons menée porte sur l'évaluation du *Stanford POS Tagger* sur l'ensemble de données P7T+. Il s'agit de connaître la performance maximale que peut obtenir le système lorsqu'il est entraîné sur des données qui ont été manuellement validées. Les résultats que nous présentons ici ont été obtenus en validation croisée en 10 strates. L'écart type (σ) des scores calculés sur les différentes strates est également reporté. Les résultats sont présentés dans la table 1. Le *Stanford POS Tagger* obtient une précision moyenne de 96,93% sur les tokens et de 50,03% sur les phrases. Ces résultats sont conformes à l'état-de-l'art des méthodes n'utilisant pas de ressources externes (Crabbé et Candito, 2008). Il faut cependant noter que les scores présentés ne sont pas directement comparables aux approches précédentes qui n'utilisaient pas une méthodologie d'évaluation en validation croisée.

	Précision	**Min. - Max.**	**Écart type**
Tokens	96,93	96,55 - 97,28	0,219
Phrases	50,03	47,08 - 52,41	1,888
Mots inconnus	85,44	82,04 - 87,67	1,661

TABLE 1 – Scores de précision sur les tokens, phrases et mots inconnus du *Stanford POS tagger* calculés à partir du P7T+ en validation croisée en 10 strates. Le minimum, le maximum et l'écart type des scores calculés sur les 10 strates sont également reportés.

3.2 Entraînement à partir de données automatiquement annotées

Dans une seconde série d'expériences, nous évaluons la performance d'une méthode d'étiquetage morpho-syntaxique entraînée à partir de données automatiquement annotées. Les résultats sont présentés dans la table 2. Le modèle entraîné sur la totalité de *Wikinews*$^{\text{pos}}$ obtient les meilleurs scores avec une précision moyenne de 96,97% sur les tokens et de 49,74% sur les phrases. Il s'agit d'un niveau de performance statistiquement comparable [23] à celui obtenu avec le modèle entraîné sur le P7T+ (décrit à la section 3.1). Ce résultat montre qu'il est possible, compte tenu de la taille des données manuellement annotées disponibles en français à ce jour, de créer un modèle d'étiquetage morpho-syntaxique tout aussi performant à partir de données automatiquement annotées.

23. $\rho > 0{,}1$ avec un t-test de Student.

Entraînement	Préc. tokens	Préc. phrases	Préc. inconnus
Wikinews^{pos} (1:1 P7T+)	96,46	44,42	**80,81**
Wikinews^{pos} (2:1 P7T+)	96,77	47,35	80,08
Wikinews^{pos} (3:1 P7T+)	96,88	48,52	79,20
Wikinews^{pos} (4:1 P7T+)	**96,97**†	**49,57**†	78,20

TABLE 2 – Scores de précision sur les tokens, phrases et mots inconnus du *Stanford POS tagger* entraîné à partir de *Wikinews* (annoté automatiquement) et évalué sur le P7T+. Le ratio entre la taille de l'ensemble d'entraînement et la taille du P7T+ est indiqué entre parenthèses. Les scores indiqués par le caractère † n'ont pas de différence statistiquement significative par rapport aux scores obtenus par le modèle entraîné sur le P7T+ ($\rho > 0{,}1$ avec un t-test de Student).

Il est intéressant de voir que la précision sur les tokens et les phrases est en constante augmentation par rapport à la taille du corpus d'entraînement et ce, malgré un nombre d'erreurs d'étiquetage automatique obligatoirement à la hausse. La précision moyenne sur les mots inconnus est quant à elle en diminution. Néanmoins, le nombre total d'erreurs commises sur les mots inconnus est en nette diminution (7128 mots inconnus mal étiquetés avec le modèle entraîné à partir d'un facteur 1 du P7T+ contre 5168 avec le modèle entraîné sur 100% *Wikinews*^{pos}). On peut également constater qu'il faut une quantité bien plus importante de données automatiquement annotées que de données manuellement annotées, ici quatre fois plus, pour obtenir le même niveau de performance.

Entraînement	Préc. tokens	Préc. phrases	Préc. inconnus
Europarl^{pos} (1:1 P7T+)	95,85	40,22	79,45
Europarl^{pos} (4:1 P7T+)	96,53	45,51	77,46
Europarl^{pos} (8:1 P7T+)	96,74	47,38	76,68
Europarl^{pos} (16:1 P7T+)	**96,93**†	**49,22**‡	75,81
Sequoia	93,99	28,42	83,49

TABLE 3 – Scores de précision sur les tokens, phrases et mots inconnus du *Stanford POS tagger* entraîné à partir de *Europarl* (annoté automatiquement) et Sequoia (validé manuellement) et évalué sur le P7T+. Le ratio entre la taille de l'ensemble d'entraînement et la taille du FTB+ est indiqué entre parenthèses. Les scores indiqués par le caractère † ($\rho > 0{,}1$ avec un t-test de Student) et ‡ ($\rho > 0{,}05$ avec un t-test de Student) n'ont pas de différence statistiquement significative par rapport aux scores obtenus par le modèle entraîné sur le P7T+.

La table 3 rapporte les résultats que nous obtenons avec le corpus *Europarl*^{pos}. De par la différence de genre, il était attendu que les scores obtenus avec ce corpus soient moins élevés que ceux obtenus avec *Wikinews*^{pos}. On note que, en comparaison avec *Wikinews*^{pos}, il faut davantage de données de *Europarl*^{pos} pour obtenir un niveau de performance acceptable. Plus exactement, il semble falloir quatre fois plus de données pour obtenir les mêmes performances. Ainsi avec 16 fois plus de données que le P7T+, on arrive à une performance significative similaire à un système état-de-l'art entraîné sur celui-ci. Malgré des scores de précisions moins élevés, on observe les mêmes tendances de progression quels que soient les scores. Bien que la précision sur les mots inconnus diminue, le nombre de mots inconnus mal étiquetés est également à la baisse.

Dans la même table, nous présentons à titre de comparaison les résultats obtenus par un modèle entraîné sur Sequoia, seul corpus librement disponible à ce jour. Les scores de précision de ce modèle évalué sur le P7T+ sont bien en dessous de ceux obtenus par les modèles entraînés sur Wikinews^{pos} et *Europarl*^{pos}, avec une précision de 93,99% sur les tokens et de seulement 28,42% sur les phrases. Ces résultats confirment qu'un ensemble de données automatiquement annotées représente une alternative pertinente pour l'entraînement de modèles d'étiquetage morpho-syntaxique.

3.3 Performance sur un corpus sans lien de filiation

La troisième et dernière expérience que nous avons menée consiste à évaluer la performance des modèles entraînés à partir de *Wikinews*^{pos} et du P7T+ sur un corpus autre que le *French TreeBank*. Pour cela nous avons choisi le corpus Sequoia. Ce dernier est composé de phrases provenant de quatre origines : Europarl français, le journal l'Est Républicain, Wikipedia Fr et des documents de l'Agence Européenne du Médicament (EMEA). Les résultats sont présentés dans la table 4.

D'une manière générale, les scores de précisions sont plus faibles que ceux observés sur le P7T+. La taille très restreinte de Sequoia (3204 phrases) ne permet cependant pas d'établir des conclusions. Les meilleurs scores sont obtenus sur les phrases provenant de l'Est Républicain et les moins bons sur celles provenant de documents de l'EMEA (domaine médical). Il s'agit d'un comportement normal puisque les modèles ont été construits à partir de phrases issues de documents journalistiques. Encore une fois, les résultats du modèle entraîné sur *Wikinews*^{pos} sont très proches de ceux obtenus par le modèle entraîné sur le P7T+.

Entraînement	Europarl	Est Rép.	Wikipedia	EMEA	Tout
FTB+	94,00	**95,10**	94,86	92,06	93,85
Wikinews^{pos}	93,55	**94,56**	94,61	91,09	93,30

TABLE 4 – Scores de précision sur les tokens du *Stanford POS tagger* entraîné à partir de *Wikinews*^{pos} et du FTB+ et évalué sur le Sequoia. Les scores de précision en fonction de l'origine des phrases sont également reportés.

4 Travaux connexes relatifs à la construction de corpus

La procédure d'annotation morpho-syntaxique de corpus repose en général sur une procédure en deux étapes [24] : d'abord une assignation automatique des étiquettes par un étiqueteur existant (étape aussi appelée «pré-annotation») et ensuite une révision de celles-ci par des annotateurs humains (Hajičová et al., 2010). On retrouve cette manière de précéder dans la construction des corpus *Penn Treebank* (Marcus et al., 1993), *PAROLE*, *MULTEXT JOC* (Véronis et Khouri, 1995), *French Treebank* (Abeillé et al., 2003), *FREEBANK* (Salmon-Alt et al., 2004), *TCOF-POS* (un corpus libre de français parlé) (Benzitoun et al., 2012) et *Sequoia* (Candito et Seddah, 2012).

24. Le processus de construction d'un corpus annoté est plus complexe et comprend notamment les étapes suivantes : sélection et constitution de la base de textes à annoter, définition du schéma d'annotation, mise en place du protocole de validation par les experts, entraînement et mesure du taux d'accord entre ceux-ci.

Cette phase de post-édition, connue comme étant toujours nécessaire, constitue une entreprise coûteuse en temps et pécuniairement. (Fort et Sagot, 2010) montrent néanmoins qu'il suffit d'un petit corpus d'entraînement pour construire un système produisant une pré-annotation de qualité suffisante pour permettre une annotation par correction plus rapide qu'une annotation manuelle. Dans ce travail, nous ne nous situons pas dans une perspective d'un post-traitement correctif manuel.

Différentes techniques ont été proposées pour rendre plus fiable l'assignation automatique d'étiquettes ainsi que pour faciliter le travail des annotateurs en détectant (voire en corrigeant) les erreurs d'annotation. En ce qui concerne l'assignation automatique, (Clark et al., 2003) utilisent deux étiqueteurs morpho-syntaxiques pour annoter de nouvelles données et étendre leur corpus d'entraînement avec une sélection de celles-ci. Leur idée consiste à sélectionner les phrases qui maximisent l'accord d'annotation entre les étiqueteurs et d'ajouter celles-ci aux données d'entraînement, puis de recommencer la procédure. Les auteurs constatent que le co-entraînement permet d'améliorer la performance des systèmes entraînés à partir d'une quantité de données manuellement annotée très faible. Cette approche trouve son utilité lorsque l'on dispose de peu de quantité de données annotés pour entraîner un système.

L'idée de combiner plusieurs étiqueteurs se retrouve dans d'autres travaux. (Loftsson et al., 2010), par exemple, entraînent cinq étiqueteurs sur un même corpus (le corpus *Icelandic Frequency Dictionary* (IFD)), et utilisent leur combinaison pour annoter un second corpus. La combinaison [25] se fait par vote à la majorité et par degré de confiance dans les étiqueteurs en cas d'égalité. Le résultat de cette combinaison est ensuite sujet à la détection d'erreurs en utilisant la détection d'incohérences entre un étiquetage en constituants fourni par un outil tiers et l'étiquetage morpho-syntaxique des mots contenus dans les constituants (Loftsson, 2009). La correction effective des erreurs est ensuite réalisée manuellement. Les auteurs montrent que la combinaison des étiqueteurs permet d'augmenter la précision de l'étiquetage comparativement aux performances individuelles de chacun des étiqueteurs. La raison invoquée pour expliquer le phénomène est que les différents étiqueteurs produisent différentes erreurs et que cette différence peut souvent être exploitée pour conduire à de meilleurs résultats.

Sur le français, le travail qui se rapproche le plus de ces efforts est celui de (Dejean et al., 2010) pour qui le développement d'un corpus annoté morpho-syntaxiquement reste avant tout un moyen d'atteindre leur objectif : construire un étiqueteur morpho-syntaxique libre du français. Les auteurs observent (après alignement des jeux d'étiquettes) les divergences d'annotations des étiqueteurs de (Brill, 1994) (Brill) et de (Schmid, 1994) (TreeTagger). Ces observations les conduisent à émettre des règles correctives sur le résultat de la combinaison de ces étiqueteurs, qu'ils utilisent pour entraîner un étiqueteur état-de-l'art. Leurs expérimentations sont réalisées sur un corpus de près de 500 000 mots construit à partir d'extraits de Wikipédia, Wikiversity et Wikinews. L'étiqueteur est entraîné sur une partie du corpus et ses résultats sont comparés sur une autre partie par rapport aux sorties produites par l'étiqueteur Brill. Le fait que la mise au point des étiqueteurs Brill et TreeTagger n'aient pas été réalisée sur un même corpus ainsi que l'absence de corpus de référence pour évaluer les étiquetages produits, rendent difficile l'interprétation de ces résultats.

Afin d'assister la tâche de correction de corpus annotés, (Dickinson et Meurers, 2003) proposent, dans le cadre du projet DECCA [26], de s'appuyer sur l'observation des variations d'annotations

25. http://combitagger.sourceforge.net
26. http://decca.osu.edu

associées à un même *n*-gramme de mots pour trouver des erreurs d'étiquetage. L'hypothèse qu'ils font est qu'un mot ambigu peut avoir différentes étiquettes dans différents contextes mais plus ses contextes d'occurrences sont similaires, plus rare devrait être la variation d'étiquetage ; et par conséquent plus grande devrait être la probabilité qu'il s'agisse d'une erreur. Appliqué sur le corpus du Wall Street Journal (WSJ), il observe que 97,6% des variations ramenées pour des *n*-grammes de taille supérieure à 6 constituent des erreurs effectives.

Poursuivant le même objectif, (Loftsson, 2009) s'appuie sur cette technique ainsi que sur deux autres : le vote de plusieurs étiqueteurs automatiques et la cohérence de l'étiquetage morpho-syntaxique des mots en regard d'une analyse en constituants des phrases. Il observe que ces techniques permettent individuellement de détecter des erreurs et qu'elles agissent en complémentarité ; ce qui lui permet de corriger manuellement 0,23% (1 334 tokens mots) du corpus IFD. Nous notons que les deux premières techniques ne sont pas dépendantes de la langue mais que la dernière repose sur l'écriture de règles ad'hoc issues de l'observation des données.

(Boudin et Hernandez, 2012) appliquent sur le P7T des techniques de détection d'erreurs fondées sur les travaux de (Dickinson et Meurers, 2003) ainsi que des heuristiques pour assigner automatiquement des étiquettes morpho-syntaxiques aux mots composants. Ils montrent que ces corrections améliorent les performances de systèmes d'étiquetage état-de-l'art.

5 Conclusion et perspectives

Dans cet article nous montrons qu'à partir d'une certaine quantité de données pré-annotées automatiquement il est possible d'entraîner des étiqueteurs morpho-syntaxiques qui produisent des résultats équivalents à des systèmes entraînés sur des données validées manuellement. La conséquence directe de ce résultat découle de la nature des données utilisées pour ces expériences (à savoir *Wikinews* et *Europarl*) : il est possible de construire un corpus libre annoté morpho-syntaxiquement offrant une modernisation perpétuelle des textes et qui puisse servir de base pour entraîner des étiqueteurs morpho-syntaxiques statistiques produisant des analyses état-de-l'art.

Les perspectives à ce travail sont triples : d'abord confirmer la qualité de l'étiquetage automatique des annotations morpho-syntaxiques du corpus ainsi construit, ensuite étendre les annotations du corpus à d'autres niveaux d'analyse, et enfin diffuser librement la ressource par un moyen qui permette un enrichissement collaboratif. Concernant l'amélioration de la qualité d'étiquetage, (Schluter et van Genabith, 2007; Loftsson et al., 2010; Boudin et Hernandez, 2012) ont montré des pistes pour la détection et la correction d'erreurs par des procédures automatiques en utilisant la détection de variations d'étiquetage ou la combinaison de multiples étiqueteurs. Concernant l'extension du corpus à d'autres niveaux d'analyses, (Candito et Seddah, 2012) utilisent pour le projet Sequoia différentes techniques pour pré-annoter automatiquement le niveau syntaxique avec des analyses en constituants et en dépendances. Les solutions mises en oeuvre dans le projet DECCA (cf. note 26) permettent d'envisager la détection d'erreurs à ces niveaux. L'une des difficultés sera de voir s'il est possible d'automatiser certaines corrections comme dans (Boudin et Hernandez, 2012) ainsi que de voir si la taille des données annotées a une incidence sur la qualité des systèmes entraînés. L'enjeu de la mise au point de telles techniques est énorme puisqu'il s'agit de pouvoir offrir à la communauté un large corpus annoté croissant continuellement sous une licence d'exploitation offrant à l'utilisateur le droit de copier, modifier et utiliser la ressource pour la finalité qu'il souhaite.

Références

ABEILLÉ, A., CLÉMENT, L. et TOUSSENEL, F. (2003). Building and using Parsed Corpora, chapitre Building a treebank for French. Language and Speech series, Kluwer, Dordrecht.

ARUN, A. et KELLER, F. (2005). Lexicalization in crosslinguistic probabilistic parsing : The case of French. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), pages 306–313, Ann Arbor, Michigan.

BENOÎT, S. et BOULLIER, P. (2008). Sxpipe 2 : architecture pour le traitement pré-syntaxique de corpus bruts. Traitement Automatique des Langues, 49(2):155–188.

BENZITOUN, C., FORT, K. et SAGOT, B. (2012). TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe. In Actes de la conférence conjointe JEP-TALN-RECITAL, pages 99–112, Grenoble, France. Quaero.

BOUDIN, F. et HERNANDEZ, N. (2012). Détection et correction automatique d'erreurs d'annotation morpho-syntaxique du french treebank. In Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN, pages 281–291, Grenoble, France. ATALA/AFCP.

BRILL, E. (1994). Some advances in rule-based part of speech tagging. In Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI), pages 722–727.

CANDITO, M. et SEDDAH, D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In 19e conférence sur le Traitement Automatique des Langues Naturelles, Grenoble, France.

CANDITO, M.-H., CRABBÉ, B. et DENIS, P. (2010a). Statistical french dependency parsing : Treebank conversion and first results. In Proceedings of LREC, Valletta, Malta.

CANDITO, M.-H., NIVRE, J., DENIS, P. et ANGUIANO, E. H. (2010b). Benchmarking of statistical dependency parsers for french. In COLING'2010 (poster session), Beijing, China.

CLARK, S., CURRAN, J. et OSBORNE, M. (2003). Bootstrapping pos-taggers using unlabelled data. In DAELEMANS, W. et OSBORNE, M., éditeurs : Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, pages 49–55.

CONSTANT, M., TELLIER, I., DUCHIER, D., DUPONT, Y., SIGOGNE, A. et BILLOT, S. (2011). Intégrer des connaissances linguistiques dans un CRF : application à l'apprentissage d'un segmenteur-étiqueteur du français. In Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2011), Montpellier, France.

CRABBÉ, B. et CANDITO, M. (2008). Expériences d'analyse syntaxique statistique du français. In Actes de la 15ème conférence sur le Traitement Automatique des Langues Naturelles (TALN), Avignon, France.

DANLOS, L., ANTOLINOS-BASSO, D., BRAUD, C. et ROZE, C. (2012). Vers le FDTB : French Discourse Tree Bank. In Actes de la 19e conférence sur le Traitement Automatique des Langues Naturelles (TALN), pages 471–478, Grenoble, France.

DEJEAN, C., FORTUN, M., MASSOT, C., POTTIER, V., POULARD, F. et VERNIER, M. (2010). Un étiqueteur de rôles grammaticaux libre pour le français intégré à Apache UIMA. In Actes de la 17e Conférence sur le Traitement Automatique des Langues Naturelles, Montréal, Canada.

DENIS, P. et SAGOT, B. (2010). Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morpho-syntaxique état-de-l'art du français. In Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2010), Montréal, Canada.

DICKINSON, M. et MEURERS, W. D. (2003). Detecting errors in part-of-speech annotation. In Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03), pages 107–114, Budapest, Hungary.

FORT, K. et SAGOT, B. (2010). Influence of pre-annotation on pos-tagged corpus development. In Proceedings of the Fourth Linguistic Annotation Workshop, pages 56–63, Uppsala, Sweden. Association for Computational Linguistics.

GABBARD, R., MARCUS, M. et KULICK, S. (2006). Fully parsing the penn treebank. In Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, HLT-NAACL '06, pages 184–191, Stroudsburg, PA, USA. Association for Computational Linguistics.

GREEN, S., de MARNEFFE, M.-C., BAUER, J. et MANNING, C. D. (2011). Multiword expression identification with tree substitution grammars : A parsing tour de force with french. In EMNLP.

HAJIČOVÁ, E., ABEILLÉ, A., HAJIČ, J., MIROVSKÝ, J. et UREŠOVÁ, Z. (2010). Handbook of Natural Language Processing, chapitre Treebank Annotation. Chapman & Hall/CRC.

KOEHN, P. (2005). Europarl : A parallel corpus for statistical machine translation. In MT Summit.

LOFTSSON, H. (2009). Correcting a POS-tagged corpus using three complementary methods. In Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), pages 523–531, Athens, Greece. Association for Computational Linguistics.

LOFTSSON, H., YNGVASON, J. H., HELGADÓTTIR, S. et RÖGNVALDSSON, E. (2010). Developing a pos-tagged corpus using existing tools. In Proceedings of LREC.

MARCUS, M. P., MARCINKIEWICZ, M. A. et SANTORINI, B. (1993). Building a large annotated corpus of english : the penn treebank. Computational Linguistics, 19(2):313–330.

NASR, A., BÉCHET, F. et REY, J.-F. (2010). Macaon : Une chaîne linguistique pour le traitement de graphes de mots. In Traitement Automatique des Langues Naturelles - session de démonstrations, Montréal.

NIELSEN, M. (2011). Reinventing Discovery : The New Era of Networked Science. Princeton, N.J. Princeton University Press.

SAGOT, B., RICHARD, M. et STERN, R. (2012). Annotation référentielle du Corpus Arboré de Paris 7 en entités nommées. In Actes de la 19e conférence sur le Traitement Automatique des Langues Naturelles (TALN), pages 535–542, Grenoble, France.

SALMON-ALT, S., BICK, E., ROMARY, L. et PIERREL, J.-M. (2004). La FReeBank : vers une base libre de corpus annotés. In Traitement Automatique des Langues Naturelles - TALN'04, Fès, Maroc.

SCHLUTER, N. et van GENABITH, J. (2007). Preparing, restructuring, and augmenting a french treebank : lexicalised parsers or coherent treebanks ? In Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING), Melbourne, Australia.

SCHMID, H. (1994). Probabilistic part-of-speech tagging using decision trees. In Proceedings of the Conference on New Methods in Language Processing, Manchester, UK.

TOUTANOVA, K., KLEIN, D., MANNING, C. et SINGER, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In Proceedings of the 3rd Conference of the North American Chapter of the ACL (NAACL 2003), pages 173–180. Association for Computational Linguistics.

VÉRONIS, J. et KHOURI, L. (1995). Etiquetage grammatical multilingue : le projet multext. Traitement Automatique des Langues, 36(1/2):233–248.