Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical

Marie Candito[1] Djamé Seddah[1, 2]
(1) Alpage (Univ. Paris Diderot & INRIA), 175 rue du Chevaleret, 75013 Paris, France
(2) Univ. Paris Sorbonne, 28, rue Serpente, 75006 Paris, France
marie.candito@linguist.jussieu.fr, djame.seddah@paris-sorbonne.fr

RÉSUMÉ

Nous présentons dans cet article la méthodologie de constitution et les caractéristiques du corpus Sequoia, un corpus en français, syntaxiquement annoté d'après un schéma d'annotation très proche de celui du French Treebank (Abeillé et Barrier, 2004), et librement disponible, en constituants et en dépendances. Le corpus comporte des phrases de quatre origines : Europarl français, le journal *l'Est Républicain*, Wikipédia Fr et des documents de l'Agence Européenne du Médicament, pour un total de 3204 phrases et 69246 tokens. En outre, nous présentons une application de ce corpus : l'évaluation d'une technique d'adaptation d'analyseurs syntaxiques probabilistes à des domaines et/ou genres autres que ceux du corpus sur lequel ces analyseurs sont entraînés. Cette technique utilise des clusters de mots obtenus d'abord par regroupement morphologique à l'aide d'un lexique, puis par regroupement non supervisé, et permet une nette amélioration de l'analyse des domaines cibles (le corpus Sequoia), tout en préservant le même niveau de performance sur le domaine source (le FTB), ce qui fournit un analyseur multi-domaines, à la différence d'autres techniques d'adaptation comme le self-training.

ABSTRACT

The Sequoia corpus : syntactic annotation and use for a parser lexical domain adaptation method

We present the building methodology and the properties of the Sequoia treebank, a freely available French corpus annotated following the French Treebank guidelines (Abeillé et Barrier, 2004). The Sequoia treebank comprises 3204 sentences (69246 tokens), from the French Europarl, the regional newspaper *l'Est Républicain*, the French Wikipedia and documents from the European Medicines Agency. We then provide a method for parser domain adaptation, that makes use of unsupervised word clusters. The method improves parsing performance on target domains (the domains of the Sequoia corpus), without degrading performance on source domain (the French treenbank test set), contrary to other domain adaptation techniques such as self-training.

MOTS-CLÉS : Corpus arboré, analyse syntaxique statistique, adaptation de domaine.

KEYWORDS: Treebank, statistical parsing, parser domain adaptation.

1 Introduction

L'analyse syntaxique statistique a fait de grands progrès ces quinze dernières années, avec de très nombreux travaux, majoritairement sur l'anglais, fondés sur un apprentissage sur les sections du

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 321–334,
Grenoble, 4 au 8 juin 2012.

Wall Street Journal du Penn Treebank (Marcus *et al.*, 1993). D'autres langues ont bénéficié de ces avancées, à la condition, de taille, que soit disponible pour ces langues un corpus arboré, en constituants ou en dépendances. Cependant, les analyseurs ainsi obtenus, appris sur un corpus bien précis, ont leur performance maximale sur des textes similaires à ce corpus, mais sont peu robustes : ils montrent une qualité nettement dégradée lorqu'ils sont évalués sur des textes de domaine ou genre différents. C'est particulièrement vrai pour l'anglais, car le WSJ montre peu de variété de thèmes : (McClosky *et al.*, 2006) rapporte que l'analyseur de Charniak (Charniak, 2000) obtient une F-mesure en constituants labelés de 89.7% sur la section de test du WSJ, mais chute à 82.9% sur le corpus de test du Brown corpus (Francis et Kucera, 1964), corpus anglais de genres variés.

Pour le français, le French Treebank (ci-après FTB) (Abeillé et Barrier, 2004) a servi de corpus d'entraînement pour des analyseurs initialement développés pour l'anglais (voir (Seddah *et al.*, 2009) pour une comparaison de plusieurs analyseurs en constituants, et (Candito *et al.*, 2010b) pour une comparaison d'analyseurs en dépendances, pour le français). Le FTB est un corpus de phrases du journal *Le Monde*, annotées en morphologie et en constituants. Les évaluations disponibles des analyseurs appris sur le FTB sont dites *intra-domaine* : elles sont classiquement faites sur une partie du FTB, non vue à l'apprentissage. Les évaluations dites *hors-domaine*, c'est-à-dire simplement sur des phrases d'origine différente de celles du corpus d'apprentissage se heurtent à l'absence de corpus annotés dans le même schéma que le FTB. Le corpus EASy (Paroubek *et al.*, 2005) comprend des phrases de domaines et genres textuels divers, mais son format mixte entre constituants (chunks) et dépendances (dépendances entre chunks) rend difficile l'évaluation des performances d'un analyseur en constituants sur ces textes.

Pour cette raison, nous avons entrepris l'annotation syntaxique de quatre corpus en suivant, à quelques exceptions près, le schéma d'annotation du FTB, regroupés sous le nom de *corpus Sequoia* [1]. Nous présentons ici la méthodologie d'annotation et les caractéristiques du corpus arboré obtenu, ainsi que l'application sur ces corpus d'une méthode d'adaptation à de nouveaux domaines d'un analyseur statistique appris sur le FTB [2]. Si l'objectif premier est de pouvoir tester et améliorer la robustesse d'analyseurs statistiques, ces corpus, librement disponibles [3], sont utilisables à d'autres fins, en particulier pour des études linguistiques.

Nous décrivons section 2 la méthodologie d'annotation et les caractéristiques du corpus, puis section 3 la méthode d'adaptation d'analyseur et les travaux antérieurs dans ce domaine, et en section 4 les expériences réalisées et les résultats obtenus. Enfin nous concluons en section 5.

2 Les corpus Sequoia

2.1 Origine et méthode de sélection

Le corpus Sequoia comporte des phrases ou textes de quatre origines différentes : l'agence européenne du médicament, Europarl, le journal régional *l'Est Républicain* et Wikipedia Fr. Le choix de ces quatre origines est en partie conjoncturel, car lié à la disponibilité des corpus : nous avons en effet eu le souci que les corpus soient librement disponibles, et qu'ils offrent une

1. Du nom du projet (SEQUOIA ANR-08-EMER-013) ayant financé l'annotation manuelle.
2. Cet article étend un article court publié à IWPT 2011 (Candito *et al.*, 2011), relatant des expériences d'adaptation d'analyseurs sur un des quatre sous-corpus aujourd'hui disponibles.
3. https ://www.rocq.inria.fr/alpage-wiki/tiki-index.php ?page=CorpusSequoia

diversité variable par rapport au genre journalistique du FTB (diversité évaluée a priori, non précisément). D'autres critères ont guidé notre choix, comme l'existence d'autres annotations pour les phrases sélectionnées et la disponibilité de gros volume de corpus brut de même origine, en vue d'expériences d'apprentissage semi-supervisé.

2.1.1 Domaine médical

Nous avons sélectionné le domaine médical comme domaine potentiellement très éloigné de celui du FTB. Plus précisément, nous avons retenu deux documents provenant de la partie en français du corpus EMEA, lui-même inclus dans le corpus OPUS (Tiedemann, 2009) [4].

Le corpus EMEA contient des documents concernant des médicaments, essentiellement des rapports public d'évaluation (EPAR), chaque rapport étant dédié à la justification de l'autorisation ou l'interdiction de la mise sur le marché d'un médicament. La partie française que nous utilisons contient environ 1000 documents convertis d'un format pdf, et concaténés. Il s'agit pour la majorité de traductions de versions originales anglaises. D'après les procédures standards de l'Agence Européenne du médicament pour les EPARs [5] les documents sont d'abord écrits en anglais, dans des termes "compréhensibles par quelqu'un qui n'est pas expert du domaine". La traduction dans les différentes langues officielles de l'Union Européenne est gérée par le Centre de Traduction de l'UE (CdT), avec une terminologie standardisée pour la biomédecine. D'après ce que nous avons pu juger, la traduction est de très bonne qualité.

Pour l'annotation manuelle, nous avons sélectionné deux EPARs, pour constituer un corpus de développement et un corpus de test (ci-après **EMEA-dev** et **EMEA-test**). Ces deux sous-corpus sont particulièrement éloignés des phrases journalistiques, pour ce qui est du domaine (ici médical) et du genre textuel (rapport scientifique). Lexicalement, ils contiennent de la terminologie spécialisée (protocoles de test et administration de médicaments, descriptions de maladies, symptômes et contre-indications). Syntaxiquement on peut noter de nombreux impératifs (pour les instructions d'utilisation), la description de dosages, et un usage fréquent de précisions apparaissant entre parenthèses (gloses de termes savants, abréviations, information de fréquence).

2.1.2 FrWiki

La deuxième source retenue est la Wikipédia en français. Nous avons pioché dans le corpus Wikipedia Fr faisant partie du corpus PASSAGE (Villemonte De La Clergerie *et al.*, 2008) [6], le texte correspondant à 19 entrées Wikipedia, concernant des "affaires" sociales ou politiques célèbres, pour la plupart récentes. Chaque entrée correspond à une description, en général chronologique, de l'affaire en question. Ainsi nous obtenons un sous-corpus d'un genre textuel narratif, pour lequel d'autres annotations existent (PASSAGE).

2.1.3 EstRépublicain

Le corpus *L'Est Républicain* est un corpus librement disponible au CNRTL [7], rassemblant les articles de deux années de ce quotidien régional (pour un total de 150 millions de tokens ponctuation

4. opus.lingfil.uu.se/EMEA.php
5. Document 3131, sur : www.ema.europa.eu
6. Les 19 premières entrées du fichier frwiki_50.txt
7. *http ://www.cnrtl.fr/corpus/estrepublicain*

comprise). Nous avons retenu 39 articles, qui sont ceux sélectionnés dans le cadre du projet ANR ANNODIS [8] dédié à l'annotation discursive pour le français, avec comme critère d'obtenir des textes intéressants du point de vue discursif.

Avec ce choix, d'une part nous espérons qu'il sera profitable de disposer pour ce corpus à la fois des annotations syntaxiques et des annotations discursives. D'autre part, nous obtenons un sous-corpus dont le domaine est éloigné du FTB. En effet les articles retenus relatent des informations essentiellement locales (faits divers, inaugurations, ...), ce qui n'est pas le cas du FTB.

2.1.4 Europarl

Enfin, nous avons sélectionné des phrases manuellement annotées dans le cadre du projet PASSAGE (Villemonte De La Clergerie *et al.*, 2008), en choisissant une sous-partie des phrases d'Europarl sélectionnées dans le cadre de ce projet.

Trois raisons principales expliquent ce choix : (i) Europarl constitue un corpus très utilisé en TAL, un corpus arboré peut en permettre une étude fine ; (ii) le type textuel d'Europarl, débat parlementaire, montre a priori des caractéristiques syntaxiques qui peuvent différer du type journalistique, ne serait-ce que par exemple le recours fréquent à la première personne et au vocatif, et enfin (iii) les phrases choisies ont également été annotées dans le schéma d'annotation Easy (pour le projet PASSAGE), ce qui peut aider à la conversion de schémas des corpus PASSAGE, Easy vers FTB et vice-versa.

2.2 Annotation morpho-syntaxique

2.2.1 Schéma d'annotation

Choix linguistiques

Notre objectif est d'obtenir des corpus compatibles avec le FTB, et donc en suivant les choix linguistiques du FTB, caractérisé comme un schéma syntagmatique surfacique, avec des annotations fonctionnelles pour les dépendants des verbes. Ainsi avons-nous suivi autant que possible les guides d'annotation du FTB (Abeillé et Clément, 2006; Abeillé *et al.*, 2004; Abeillé, 2004).

Une exception notable concerne le traitement des mots composés. Pour les composés ni nominaux ni verbaux, nous nous sommes appuyés sur les composés existants dans le FTB. Pour les composés verbaux à syntaxe régulière, nous avons préféré n'en annoter aucun, et privilégier une analyse syntagmatique. En effet ils sont potentiellement discontinus, et leur notation est alors variable dans le FTB (par exemple, annotation *il est_en_train de...* versus *il est justement en train de ...*). Concernant les composés nominaux, le FTB contient de nombreuses incohérences (séquences de même sémantique parfois codées comme composés, parfois codées par un syntagme), en particulier pour les composés syntaxiquement réguliers à sémantique tout ou partiellement compositionnelle [9]. Nous avons donc choisi de systématiquement coder syntagmatiquement des séquences syntaxiquement régulières (comme *N prep N* ou *N A* par exemple), y compris celles pouvant être considérées comme des noms composés. Cela a le mérite de l'uniformité, mais

8. http ://w3.erss.univ-tlse2.fr/annodis

9. Par exemple *pays industrialisés* apparaît deux fois comme composé, et 41 fois comme deux mots ; *taux d'intérêt apparaît 80 comme composé, et 25 fois comme trois mots.*

appelle des traitements ultérieurs pour repérer en particulier les cas de composés sémantiquement non compositionnels.

Format

En ce qui concerne le format, au lieu de reproduire le format XML du FTB, nous avons opté pour un format certes moins riche mais beaucoup plus souple : un format parenthésé avec une ligne par phrase syntagmatiquement annotée, qui fournit la catégorie morpho-syntaxique des tokens, et leur structure syntagmatique. Ce format est celui du PennTreebank, qui s'est imposé comme format d'apprentissage des analyseurs syntagmatiques probabilistes pour diverses langues et c'est sous cette forme que nous utilisons le FTB dans nos expériences d'analyse syntaxique probabiliste.

Voici un exemple dans le format en constituants parenthésé, provenant du corpus médical :

((SENT (PP-MOD (P Afin_de) (VPinf (VN (VINF diminuer)) (NP-OBJ (DET le) (NC risque) (PP (P de) (NP (ADJ faibles) (NC valeurs) (PP (P d') (NP (NC ACT)))))))) (PONCT ,) (NP-SUJ (DET le) (NC produit) (VPpart (VPP reconstitué) (COORD (CC et) (VPpart (VPP dilué))))) (VN (V doit)) (VPinf-OBJ (VN (VINF être) (ADV bien) (VPP mélangé))) (COORD (CC puis) (VN (V doit)) (VPinf (VN (VINF être) (VPP administré)) (PP-MOD (P en) (NP (NC bolus))) (PP-MOD (P par) (NP (NC poussée) (AP (ADJ intraveineuse)) (AP (ADJ rapide)))))) (PONCT .)))

Le jeu de catégories morpho-syntaxiques que nous utilisons est celui mis ou point par (Crabbé et Candito, 2008), contenant 28 catégories, qui correspondent aux combinaisons entre une des 13 catégories grossières du FTB et des informations codées dans le FTB sous forme de traits (essentiellement distinction nom commun, nom propre, mode du verbe). Il y a appauvrissement des annotations par rapport au FTB, pour ce qui est des informations morphologiques. En effet, si une partie de celles disponibles dans le FTB est encodée dans l'étiquette morpho-syntaxique, d'autres comme le lemme, le genre et le nombre ne sont pas représentés. En outre, les catégories des composants de composés n'ont pas été explicitées (un composé est directement codé comme un seul token, avec ses composants séparés par '_'). Cet appauvrissement relatif est compensé par la souplesse d'utilisation de ce format, et la disponibilité d'outils de visualisation et validation, ce qui favorise clairement la qualité des annotations, par rapport à une validation faite directement sur format XML. D'autre part, comme indiqué supra, c'est ce format parenthésé qui est utilisé pour l'analyse syntaxique probabiliste.

Conversion en dépendances

Le corpus annoté en constituants a été automatiquement converti en dépendances en utilisant le convertisseur développé pour la conversion automatique du FTB (Candito *et al.*, 2010a). Au final, le corpus Sequoia est donc disponible sous deux formes : un format parenthésé annoté en constituants [10] décoré de fonctions syntaxiques, et un format tabulé CoNLL [11] pour la version en dépendances labelées.

2.2.2 Méthodologie d'annotation

Pour obtenir le corpus Sequoia, nous avons procédé en alternant traitements automatiques et validation de ces traitements pour passer à l'étape suivante. A toutes les étapes (segmentation, tagging, parsing, annotations des fonctions), les annotations précédentes pouvaient être remises

10. Plus précisément, deux formats en constituants sont disponibles : le format standard FTB, et un format avec une représentation modifiée des infinitives introduites par des prépositions, et un syntagme supplémentaire dans les complétives, tel que décrit dans (Candito et Crabbé, 2009). Ces modifications facilitent la conversion en dépendances. La conversion de l'un vers l'autre format est automatique.

11. `http://ilk.uvt.nl/conll/#dataformat`

en cause. La séquence a été la suivante :

- Prétraitements automatiques : Segmentation en phrases, reconnaissance hors contexte de composés et tokenisation via l'outil Bonsai [12]
- Etiquetage morpho-syntaxique en utilisant le tagger MElt (Denis et Sagot, 2009)
- Validation manuelle en éditeur simple, par un seul annotateur expert, du tagging, de la segmentation en phrases, et de la reconnaissance de composés
- Pour tous les sous-corpus sauf EMEA : Analyse syntagmatique automatique au moyen de deux parsers statistiques différents, en guidant les analyseurs avec les tags manuellement validés : les analyses doivent se conformer aux catégories fournies en entrée. Les analyseurs sont le parser de Berkeley (Petrov et Klein, 2007) et l'analyseur de Charniak (Charniak, 2000), tous deux adaptés et entraînés sur le FTB. Pour EMEA : la validation syntaxique a été faite par un annotateur expert.
- Validation manuelle indépendante des deux sorties d'analyseurs, via l'outil graphique WordFreak (Morton et LaCivita, 2003) adapté pour le tagset et le jeu de fonctions du FTB, puis adjudication.
- Annotation automatique des fonctions des dépendants des verbes finis, en utilisant l'annotateur en fonctions intégré à Bonsai
- Validation manuelle des annotations fonctionnelles par deux annotateurs indépendamment, via WordFreak, puis adjudiction.
- Vérifications systématiques par un expert de points repérés comme difficiles [13] ; vérification systématique de la cohérence du traitement des composés.

2.2.3 Evaluation de l'annotation

Pour évaluer l'accord inter-annotateurs, et la distance au corpus de référence après adjudication et vérifications systématiques, nous utilisons l'outil Evalb servant habituellement à l'évaluation des sorties d'un analyseur par rapport à des analyses de référence. Pour les sous-corpus Europarl, EstRépublicain et FrWiki, nous fournissons table 1 l'accord deux à deux entre trois résultats d'annotations : l'annotation A, l'annotation B et le résultat de l'adjudication de A et B plus vérification. La mesure utilisée est la moyenne harmonique (F-mesure) entre la précision et le rappel en constituants labelés. Nous avons dû contourner le problème de tokenisations divergentes, où une séquence de tokens analysée comme un mot composé dans un des fichiers et pas dans l'autre. Par exemple la séquence *en fait* peut avoir été codée (ADV en_fait) d'un côté et (CLO en) (V fait) de l'autre. Pour résoudre ce problème, nous transformons les annotations avant l'évaluation de l'accord : tous les composés sont transformés en structure contenant les composants, avec une catégorie unique pour les composants. Pour notre exemple '(ADV en_fait)' est transformé en (ADV (Z en) (Z fait)). Ainsi les divergences de tokenisation non seulement ne bloquent pas evalb, mais sont en outre prises en compte dans l'évaluation.

L'évaluation montre des résultats assez satisfaisants pour Europarl et EstRépu, avec une nette amélioration lors de l'évaluation avec la référence. Pour FrWiki, l'accord entre les deux annotations simples est bas : il est comparable avec les résultats obtenus par l'analyseur sur le domaine neutre (section 4). C'est en effet par ce corpus que l'annotation a commencé. On voit ici que la phase de formation est longue. Sachant cela, la vérification pour ce corpus a été plus poussée.

12. http ://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html

13. Entre autres : les clivées versus relatives, le causatif, les complétives en *de* objet direct versus oblique (de-obj), le repérage d'incohérences comme par exemple des verbes finis sans sujet.

	Annotations A vs. B	Annotation A vs. référence	Annotation B vs. référence
FrWiki	83.96	91.59	88.64
Europarl	90.14	94.20	92.26
EstRépu	90.45	94.22	93.72

TABLE 1 – Evaluation deux à deux (moyenne des F-mesures) des annotations simple A, simple B et de la référence (après adjudication de A et B et vérifications systématiques).

2.3 Caractéristiques

La table 2 fournit les caractéristiques des différents corpus annotés, en regard de celle des corpus de développement et d'entraînement du FTB (FTB-dev et FTB-train) utilisés pour les expériences (section 4) [14].

	Corpus Sequoia					**FTB**	
	Médical		Neutre				
	EMEA dev	**EMEA test**	**Est Rép.**	**Euro Parl**	**Fr Wiki**	**dev**	**train**
Nb de phrases	574	544	529	561	996	1235	9881
Longueur moyenne	16,3	22,0	21,0	26,3	22,2	29,6	28,1
Ecart type sur la longueur	14,7	15,0	12,9	15,0	18,0	16,0	16,5
Données pour tout type de formes fléchies (ponctuation y compris)							
Taille du vocabulaire	1916	1737	3337	3300	4687	7222	24110
% d'inconnus	41.4	35.8	29,2	20,6	34,2	22,5	-
Nb d'occ.	9343	11964	11114	14745	22080	36508	278083
% d'occ. d'inconnus	23.0	19.7	11,2	6,6	12,9	5,2	-
% d'occ. de Noms propres	1,7	2.7	5,1	2,9	9,7	4,1	4,0
Données pour les formes alphanumériques minusculisés							
Taille du vocabulaire	1695	1599	3173	3165	4410	6904	22526
% d'inconnus	36.6	34.0	28,0	20,1	32,6	21,6	-
Nb d'occ.	8107	10451	9552	13073	18619	30940	235105
% d'occ. d'inconnus	23.2	20.9	12,1	7,0	13,8	5,7	-

TABLE 2 – Caractéristiques chiffrées des corpus manuellement annotés. Les *inconnus* sont les formes absentes du FTB-train.

Les différents nouveaux corpus ont chacun environ 500 phrases, sauf FrWiki (961 phrases). Si la longueur moyenne des phrases varie nettement entre les différents corpus, on peut noter une grande variance. Ce sont les phrases du FTB qui sont les plus longues en moyenne (29,6 pour

14. Pour comparabilité avec nos résultats antérieurs, nous utilisons la partie du FTB annotée en fonctions grammaticales, telle que distribuée en 2007, qui contient 12351 phrases. La version actuellement disponible du FTB contient environ 4000 phrases supplémentaires. Nous utilisons le découpage initialement proposé par (Crabbé et Candito, 2008) en corpus de test (1235 premières phrases), corpus de développement (1235 phrases suivantes) et 9881 phrases restantes comme corpus d'apprentissage. Le corpus original XML est prétraité tel que décrit dans (Candito et Crabbé, 2009). En particulier les composés nominaux et verbaux syntaxiquement réguliers sont défaits et représentés syntagmatiquement, et chaque occurrence de composé restante traitée comme un seul token (par exemple *(N (P à) (N cause) (P de))* est remplacé par *(N à_cause_de)*).

FTB-dev et 28,1 pour FTB-train), devant même Europarl (26,3).
La table fournit également la taille des vocabulaires (de formes fléchies), et en leur sein la proportion de formes qui sont absentes du FTB-train. Nous fournissons les chiffres calculés en utilisant tous les tokens (y compris la ponctuation) ainsi que ceux calculés sur les tokens alphanumériques minusculisés [15], pour mieux évaluer la diversité lexicale. On peut constater que le corpus médical comporte de loin le vocabulaire le plus éloigné de celui du FTB (plus d'une forme sur trois est absente du FTB-train). Pour le corpus EMEA-dev, la proportion d'inconnus en comptant tous les types de formes fléchies est très haute, du fait d'un grand nombre de mots entièrement capitalisés (la proportion passe de 41,4 à 36,6 en ignorant la ponctuation et en minusculisant). Pour le corpus FrWiki, la forte proportion d'inconnus (34,2%) peut s'expliquer par une grande fréquence des noms propres (cf. la ligne *% d'occurrences de noms propres* : environ une occurrence sur 10 est un nom propre dans FrWiki).

Les lignes sur les nombres d'occurrences et le pourcentage d'inconnus parmi ces occurrences donnent une vision plus précise de la diversité lexicale des corpus. Dans les corpus médicaux, une occurrence sur 5 (et presque une sur 4 pour EMEA-dev) correspond à un inconnu du FTB-train, ce qui, avec la faible proportion d'occurrences de noms propres (1,7 et 2,7) indique que les mots inconnus sont plutôt des mots fréquemment utilisés dans ces corpus. Au contraire, pour FrWiki on voit que, calculée sur les occurrences, la proportion d'inconnus tombe à 12,9 (la majorité des inconnus du vocabulaire sont des noms propres, apparaissant rarement). Le corpus le plus proche lexicalement du FTB semble être Europarl : seulement 6,6% des occurrences sont des inconnus, formant un cinquième du vocabulaire, ce qui constitue moins d'occurrences d'inconnus que dans le FTB-dev.

3 Adaptation de domaine par pont lexical

Notre objectif est d'explorer une méthode d'amélioration des performances d'un analyseur statistique sur des textes d'origine différente de celle du corpus d'entraînement de l'analyseur, les différences pouvant relever du domaine et/ou du genre des textes. Pour simplifier, nous utilisons par la suite les termes *domaine source* pour les caractéristiques (domaine, genre, registre) du corpus d'entraînement, *domaines cibles* pour celles des textes d'origine différente et *analyse hors-domaine* pour l'analyse de textes des domaines cibles.

Pour améliorer l'analyse hors-domaine, nous proposons d'adapter une technique testée au départ pour le parsing intra-domaine. S'inspirant de l'utilisation par (Koo *et al.*, 2008) de clusters de mots comme traits d'un analyseur discriminatif en dépendances, (Candito et Crabbé, 2009) ont proposé une technique qui, en réduisant la dispersion des données lexicales, améliore les performances de parsing intra-domaine. Ils entraînent un analyseur statistique sur un corpus où les mots sont remplacés par des identifiants de clusters de mots, obtenus de manière non supervisée sur un corpus brut de grande taille. Le parsing se fait ensuite de la même manière, en remplaçant chaque mot par leur cluster correspondant, de manière déterministe et non contextuelle, puis en réinsérant les tokens originaux pour obtenir les sorties d'analyse.

Plus précisément, le regroupement de formes fléchies en clusters se fait en deux étapes :
– Les formes fléchies sont d'abord groupées en clusters morphologiques via un lexique morphologique. Il s'agit de ramener un ensemble de formes fléchies à une forme canonique, dite forme

15. Plus précisément les tokens comportant au moins une lettre ou un chiffre, et ramenés à une forme minusculisée.

défléchie, avec comme principe de conserver exactement la même ambiguïté de catégories morpho-syntaxiques (contrairement par exemple à une lemmatisation). On veut en effet déléguer la désambiguisation de catégories à l'analyseur, et ne pas trancher par pré-traitement. Pour cela, pour une forme donnée, on récupère la liste de ses catégories recensées dans le dictionnaire, puis, tant que cette liste de catégories ne varie pas, le pluriel est ramené au singulier, le féminin au masculin, et pour les formes verbales conjuguées non ambiguës, les personnes, mode et temps verbaux sont ramenées à la deuxième personne présent pluriel (moyen rapide de trouver une forme n'introduisant pas de nouvelles ambiguïtés). Par exemple, *analysées* est ramené à *analysé*, mais *entrées* est ramené à *entrée*, de manière à conserver l'ambiguïté nom/participe. Toutes les formes finies de *augmenter* sont ramenées à *augmentez*, mais par exemple *joue* est inchangé pour préserver son ambiguïté catégorielle.
– Ensuite un algorithme de clustering non supervisé (Brown *et al.*, 1992) est appliqué sur gros corpus préalablement segmenté en phrases, tokenisé et défléchi (i.e. où les formes fléchies sont remplacées par leur forme défléchie correspondante). On obtient ainsi des clusters de formes défléchies. Il s'agit d'un algorithme hiérarchique et agglomératif, où le critère de fusion de deux clusters est la perte minimale de vraisemblance dans un modèle bigramme de séquences de clusters.

Dans cet article, nous adaptons cette technique au problème spécifique de la non robustesse des analyseurs statistiques, en utilisant des clusters de mots appris sur la concaténation de corpus du domaine source (ou proche du domaine source) et des domaines cibles. L'objectif est d'obtenir que soient groupés sous le même cluster des mots appartenant au domaine source et des mots appartenant aux domaines cibles, de façon à réaliser un pont entre les vocabulaires respectifs de ces domaines (d'où le nom d'adaptation à de nouveaux domaines par "pont lexical").

3.1 Travaux antérieurs reliés

Différentes techniques ont été proposées pour adapter des modèles d'analyse existants à de nouveaux genres :
– Adaptation au domaine via de l'auto-entraînement (*self-training*) (Bacchiani *et al.*, 2006; McClosky *et al.*, 2006; Sagae, 2010) : un analyseur entraîné sur le domaine source est utilisé pour analyser du domaine cible, et on réentraîne un analyseur sur les données validées source et les données prédites cibles. Le corpus d'entraînement ainsi obtenu, bien que bruité, capture suffisamment de régularités du domaine cible pour améliorer les performances d'analyse sur ce domaine (tout en dégradant les performances sur le domaine source) ;
– co-entraînement avec sélection d'exemples (Steedman *et al.*, 2003) : deux analyseurs sont itérativement re-entraînés sur leurs sorties respectives, les phrases du domaine cible à utiliser étant choisies de manière à minimiser les erreurs d'analyse tout en maximisant l'utilité à l'entraînement ;
– transformation de treebank et adaptation du domaine cible (Foster, 2010) ;
– adaptation méticuleuse du domaine cible à la source d'entraînement (Foster *et al.*, 2007) ;

Bien que différentes, les techniques ici évoquées sont toutes conçues pour combler la variation syntaxique et lexicale entre le domaine source et les domaines cibles. La variation lexicale est en particulier problématique dans le cas d'une langue à la morphologie plus riche que l'anglais, la flexion augmentant la dispersion des données lexicales.

4 Expériences et résultats

4.1 Clusters de mots

Pour calculer les clusters de mots nous utilisons diverses concaténations de quatre corpus, avec d'une part le corpus *L'Est Républicain* déjà cité section 2, de 150 millions de tokens, qui va jouer le rôle de corpus proche du domaine source malgré des différences manifestes concernant les sujets traités [16]. Et d'autre part, nous utilisons des tronçons de corpus de même origine que les sous-corpus Sequoia annotés : Europarl, Wikipedia Fr et domaine médical. Cela donne quatre corpus :

– **ER** : 150 millions de tokens *L'Est Républicain*
– **MED** : 12 millions de tokens du domaine médical, dont 5 millions du corpus EMEA français [17] cité section 2 et 7 millions de tokens provenant du site *doctissimo* [18].
– **EP** : la même taille, soit 12 millions de tokens, d'Europarl français,
– **FW** : et 12 millions de tokens de Wikipedia Fr

Pour le calcul des clusters, les phrases contenues dans le corpus arboré Sequoia ont été retirées.

Le corpus ER, en tant que corpus journalistique régional, est choisi comme corpus proche du FTB, malgré des différences manifestes concernant les sujets traités. La concaténation du corpus ER et du corpus MED+EP+FW va jouer le rôle de pont lexical entre le domaine source (journalistique) et les domaines cibles.

Les corpus bruts ER, MED, EP et FW sont d'abord prétraités par l'outil Bonsai (segmentés en phrases, tokenisés, et des mots composés sont reconnus hors-contexte). Puis nous appliquons le processus de défléchissement décrit section 3, pour remplacer chaque forme fléchie par sa forme défléchie équivalente. Le lexique morphologique utilisé est le L*efff* (Sagot, 2010).

Enfin nous calculons des clusters de formes défléchies [19] en utilisant l'implémentation par (Liang, 2005) de l'algorithme de (Brown *et al.*, 1992) :

– les *clusters source* sont obtenus en appliquant l'outil sur le corpus ER,
– les *clusters pont mixtes* sont obtenus sur la concaténation de ER + MED + EP + FW (soit environ 186 millions de tokens).
– les *clusters pont er-med* sont obtenus sur la concaténation de ER + MED uniquement (soit environ 162 millions de tokens), pour tester la méthode avec des clusters plus ciblés sur le vocabulaire médical.

Dans les trois cas, le nombre de clusters générés est de 1000, et les formes défléchies considérées sont celles apparaissant au moins 100 fois dans le corpus d'apprentissage [20].

16. D'après les indicateurs de la table 2, c'est plutôt Europarl qui est le plus proche en termes de vocabulaire.
17. Le corpus fait initialement environ 14 millions de tokens, mais contient énormément de formules répétitives. La suppression des phrases doublons réduit sa taille à 5 millions de tokens.
18. Il s'agit des pages médicaments et des pages du glossaire. Le texte est bien formé et proche du corpus EMEA dans les thématiques. Les phrases doublons ont été retirées.
19. Nous avons réalisé des tests en utilisant des clusters calculés sur formes fléchies (sans le processus de défléchissement), ce qui donne systématiquement des résultats moins bons qu'en utilisant les clusters sur formes défléchies.
20. Nous avons constaté lors de tests qu'un seuil plus bas, a peu d'impact sur les résultats. Un seuil de 100 réduit le vocabulaire considéré ce qui limite le temps de calcul des clusters.

4.2 Protocole et expériences

Nous réalisons ces premiers tests en analyse en constituants sans annotations fonctionnelles. Tous les traitements (entraînement d'analyseur et tests) se font donc sur des versions des corpus où les annotations fonctionnelles sont supprimées [21]. Nous utilisons l'algorithme d'apprentissage et d'analyse de PCFG avec annotations latentes (ci-après PCFG-LA) de (Petrov et Klein, 2007), et son implémentation [22], avec modèle de lissage pour les mots rares et inconnus adapté au français.

Pour cet algorithme, (Petrov, 2010) montre une variabilité des résultats selon les valeurs aléatoires choisies à l'initialisation de l'algorithme EM d'apprentissage des probabilités de règles avec annotations latentes. Aussi, nous réalisons pour chaque expérience quatre exécutions de l'apprentissage, avec quatre graines aléatoires différentes. Tous les apprentissages se font en utilisant 5 cycles de fission-fusion.

Pour l'évaluation des performances, nous utilisons l'outil Evalb, et fournissons la moyenne des F-mesures de constituants labelés (moyenne sur les quatre graines aléatoires) pour les phrases de moins de 40 mots ainsi que pour toutes les phrases.

Nous utilisons PCFG-LA pour apprendre quatre analyseurs, sur quatre versions du FTB-train (cf. section 2) différant par les symboles terminaux utilisés (les feuilles lexicales) :
– **forme fléchie** : les formes fléchies sont laissées telles quelles
– **forme défléchie** : chaque forme fléchie est remplacé par sa forme défléchie équivalente
– **cluster source** : chaque forme défléchie est ensuite remplacée par son cluster source équivalent (clusters appris sur le corpus ER)
– **cluster pont mixte** : idem mais en utilisant les clusters appris sur ER + MED + EP + FW
– **cluster pont er-med** : idem mais en utilisant les clusters appris sur ER + MED

4.3 Résultats et discussion

Nous avons réalisé des tests en comparant les résultats sur le FTB et sur le corpus Sequoia. Plus précisément, d'une part avons considéré trois "domaines" : le domaine source (FTB), un domaine très éloigné (domaine médical, corpus Emea), et un domaine que nous appelons *neutre*, regroupant les autres parties du corpus Sequoia (phrases de Wikipédia Fr, Europarl et Est Républicain). D'autre part, pour chaque domaine (source, médical et neutre) nous avons séparé corpus de test pour les tests finaux, et corpus de développement pour la phase exploratoire, de la manière suivante :
– **domaine source** : FTB-dev et FTB-train tels que décrits note 14
– **domaine médical** : EMEA-dev et EMEA-train, cf. les colonnes 2 et 3 de la table 2
– **domaine neutre** : SequoiaN-dev et SequoiaN-test obtenus en découpant en deux chacun des sous-corpus annotés FrWiki, EstRep et Europarl (colonnes 4,5 et 6 table 2). Cela donne 1043 phrases pour SequoiaN-dev et autant pour SequoiaN-test.

La table 3 fournit les résultats obtenus. Dans le cas standard, où les symboles terminaux sont simplement les formes fléchies, on observe sans surprise une nette dégradation des performances entre le domaine source (F=83.6) et le domaine médical (F=78.5). La dégradation est nettement

21. En outre, nous utilisons une instantiation des corpus où des modifications automatiques de structure ont été faites, comme décrit dans (Candito et Crabbé, 2009), ceci pour faciliter la conversion en dépendances de tous les résultats d'analyse. Les modifications introduisent des syntagmes supplémentaires pour les prépositions introduisant une infinitive et les complétives.
22. `http://code.google.com/p/berkeleyparser`

moindre pour le domaine "neutre" avec F=82.2 pour le corpus SequoiaN-test. L'apprentissage sur les phrases du journal *Le Monde* se généralise donc assez bien sur ces trois autres types de corpus (FrWiki, Europarl et Est Républicain).

Les résultats obtenus avec défléchissement (ligne 2) sont meilleurs dans toutes les configurations. On note cependant que l'incrément est moindre pour les domaines cibles que pour le domaine source (les différences restent statistiquement significatives, $p < 0.05$) [23]

Enfin, les trois dernières lignes donnent les résultats lorsque les formes sont remplacées par des clusters. La technique améliore les résultats pour le parsing du domaine source, ce qui confirme des résultats précédents. Ici nous montrons qu'elle est valable également pour les deux domaines cibles. Cela constitue donc une technique qui rend plus robuste l'analyseur, en améliorant les performances sur les domaines cibles tout en améliorant également sur le domaine source, au contraire par exemple de la technique d'auto-entraînement.

En revanche, les trois configurations qui varient selon le corpus utilisé pour le calcul des clusters offrent peu de variation dans les résultats (la plupart des différences entre ces 3 lignes ne sont pas significatives (p-value > 0.05). Ceci invalide l'hypothèse selon laquelle il serait bénéfique d'utiliser un corpus permettant de faire un pont entre le vocabulaire du domaine source et celui du domaine cible. [24]

	Toutes les phrases			Phrases de moins de 40 mots		
	Médical **EMEA-test**	Neutre **SequoiaN-test**	Source **FTB-test**	Médical **EMEA-test**	Neutre **SequoiaN-test**	Source **FTB-test**
Nombre de phrases	544	1043	1235	486	919	969
Terminaux						
formes fléchies	78.5	82.2	83.6	80.5	84.4	85.7
formes défléchies	79.0	83.1	85.0	81.0	85.0	87.4
clusters source	80.8	84.1	86.0	82.6	86.0	88.3
clusters pont mixtes	80.2	84.4	86.0	82.2	86.3	88.2
clusters pont er-med	80.7	84.1	85.9	82.8	86.1	88.2

TABLE 3 – F-mesures calculées via evalb, en ignorant la ponctuation, chacune étant moyennée sur quatre graines aléatoires différentes.

5 Conclusion

Nous avons présenté le corpus arboré Sequoia, comportant quatre sous-corpus annotés syntaxiquement en suivant le schéma du French Treebank, à quelques exceptions près. Les corpus sont librement disponibles sous forme de constituants et de dépendances.

Nous avons exploité ces corpus pour évaluer une méthode d'adaptation d'un analyseur statistique à des domaines autres que celui de son corpus d'entraînement, méthode fondée sur l'utilisation de clusters de mots, proposée dans une version préliminaire de ce travail (Candito *et al.*, 2011). Nous montrons que cette technique améliore les performances sur les domaines cibles, tout en ne dégradant pas les résultats sur le domaine source, contrairement à toutes les techniques d'adaptation de parsers statistiques à notre connaissance. En revanche, les tests réalisés en faisant

23. En utilisant l'outil `http://www.cis.upenn.edu/~dbikel/software.html#comparator`.

24. Nous contredisons ici les résultats publiés à IWPT (Candito *et al.*, 2011) où pour le corpus médical, les résultats étaient légèrement meilleurs avec les clusters pont er-med. D'une part le corpus médical a légèrement été modifié lors de la phase de vérification systématique d'erreurs d'annotation, d'autre part, il semble que cette amélioration n'était pas stable lors des tests avec différentes graines aléatoires.

varier le corpus brut sur lequel calculer les clusters ne montrent pas d'avantage clair à utiliser du texte brut du domaine cible.

Remerciements

Nous remercions chaleureusement les trois annotatrices Vanessa Combet, Catherine Moreau-Mocquay et Virginie Mouilleron pour leur travail très consciencieux. L'annotation a été financée par l'ANR (projet SEQUOIA ANR-08-EMER-013).

Références

ABEILLÉ, A. (2004). Annotation fonctionnelle, version du 1er mars 2004. `http://www.llf.cnrs.fr/Gens/Abeille`.

ABEILLÉ, A. et BARRIER, N. (2004). Enriching a french treebank. *In Proc. of LREC'04*, Lisbon, Portugal.

ABEILLÉ, A. et CLÉMENT, L. (2006). Annotation morpho-syntaxique, version du 10 nov. 2006. `http://www.llf.cnrs.fr/Gens/Abeille`.

ABEILLÉ, A., TOUSSENEL, F. et CHÉRADAME, M. (2004). Corpus le monde, annotation en constituants, guide pour les correcteurs, version du 31 mars 2004. `http://www.llf.cnrs.fr/Gens/Abeille`.

BACCHIANI, M., RILEY, M., ROARK, B. et SPROAT, R. (2006). Map adaptation of stochastic grammars. *Computer speech & language*, 20(1):41–68.

BROWN, P. F., DELLA, V. J., DESOUZA, P. V., LAI, J. C. et MERCER, R. L. (1992). Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

CANDITO, M. et CRABBÉ, B. (2009). Improving generative statistical parsing with semi-supervised word clustering. *In Proceedings of IWPT 2009*, pages 138–141, Paris, France.

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

CANDITO, M., HENESTROZA ANGUIANO, E. et SEDDAH, D. (2011). A word clustering approach to domain adaptation : Effective parsing of biomedical texts. *In Proceedings of IWPT 2011*, pages 37–42, Dublin, Ireland.

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

CHARNIAK, E. (2000). A maximum entropy inspired parser. *In Proceedings of NAACL 2000*, pages 132–139, Seattle, WA.

CRABBÉ, B. et CANDITO, M. (2008). Expériences d'analyse syntaxique statistique du français. *In Actes de TALN 2008*, pages 45–54, Avignon, France.

DENIS, P. et SAGOT, B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art pos tagging with less human effort. *In Proc. of PACLIC*, Hong Kong, China.

FOSTER, J. (2010). "cba to check the spelling" : Investigating parser performance on discussion forum posts. *In Proceedings of HLT-NAACL 2010*, pages 381–384, Los Angeles, California.

FOSTER, J., WAGNER, J., SEDDAH, D. et VAN GENABITH, J. (2007). Adapting wsj-trained parsers to the british national corpus using in-domain self-training. *In Proceedings of the Tenth IWPT*, pages 33–35.

FRANCIS, W. N. et KUCERA, H. (1964). *Manual of Information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers*. Brown University, Providence, Rhode Island.

KOO, T., CARRERAS, X. et COLLINS, M. (2008). Simple semi-supervised dependency parsing. *In Proceedings of ACL-08*, pages 595–603, Columbus, USA.

LIANG, P. (2005). Semi-supervised learning for natural language. *In MIT Master's thesis*, Cambridge, USA.

MARCUS, M., MARCINKIEWICZ, M. et SANTORINI, B. (1993). Building a large annotated corpus of english : The penn treebank. *Computational linguistics*, 19(2):313–330.

MCCLOSKY, D., CHARNIAK, E. et JOHNSON, M. (2006). Reranking and self-training for parser adaptation. *In Proceedings of COLING-ACL 2006*, pages 337–344, Sydney, Australia.

MORTON, T. et LACIVITA, J. (2003). Wordfreak : an open tool for linguistic annotation. *In Proceedings of NAACL 2003, Demonstrations*, pages 17–18.

PAROUBEK, P., POUILLOT, L.-G., ROBBA, I. et VILNAT, A. (2005). Easy : Campagne d'évaluation des analyseurs syntaxiques. *In Proceedings of TALN'05, EASy workshop : campagne d'évaluation des analyseurs syntaxiques*, Dourdan.

PETROV, S. (2010). Products of random latent variable grammars. *In Proceedings of HLT-NAACL 2010*, pages 19–27, Los Angeles, California.

PETROV, S. et KLEIN, D. (2007). Improved inference for unlexicalized parsing. *In Proceedings of HLT-NAACL 2007*, pages 404–411, Rochester, New York.

SAGAE, K. (2010). Self-training without reranking for parser domain adaptation and its impact on semantic role labeling. *In Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 37–44, Uppsala, Sweden.

SAGOT, B. (2010). The L*efff*, a freely available and large-coverage morphological and syntactic lexicon for french. *In Proceedings of LREC'10*, Valetta, Malta.

SEDDAH, D., CANDITO, M. et CRABBÉ, B. (2009). Cross parser evaluation and tagset variation : a french treebank study. *In Proceedings of IWPT 2009*, IWPT '09, pages 150–161, Stroudsburg, PA, USA.

STEEDMAN, M., HWA, R., CLARK, S., OSBORNE, M., SARKAR, A., HOCKENMAIER, J., RUHLEN, P., BAKER, S. et CRIM, J. (2003). Example selection for bootstrapping statistical parsers. *In Proceedings of the NAACL 2003*, pages 157–164.

TIEDEMANN, J. (2009). News from opus - a collection of multilingual parallel corpora with tools and interfaces. *Recent advances in natural language processing V : selected papers from RANLP 2007*, 309:237.

VILLEMONTE DE LA CLERGERIE, E., HAMON, O., MOSTEFA, D., AYACHE, C., PAROUBEK, P. et VILNAT, A. (2008). Passage : from french parser evaluation to large sized treebank. *In Proceedings of LREC'2008*.