Etiquetage morpho-syntaxique en domaine de spécialité: le domaine médical

Christelle Tiana Rabary[1] Thomas Lavergne[1,2] Aurélie Névéol[1]
(1) LIMSI-CNRS, Campus Universitaire d'Orsay, bât 508, 91405 ORSAY, FRANCE
(2) Université Paris-Sud, 91403 ORSAY, FRANCE
prenom.nom@limsi.fr

Résumé. L'étiquetage morpho-syntaxique est une tâche fondamentale du Traitement Automatique de la Langue, sur laquelle reposent souvent des traitements plus complexes tels que l'extraction d'information ou la traduction automatique. L'étiquetage en domaine de spécialité est limité par la disponibilité d'outils et de corpus annotés spécifiques au domaine. Dans cet article, nous présentons le développement d'un corpus clinique du français annoté morpho-syntaxiquement à l'aide d'un jeu d'étiquettes issus des guides d'annotation French Treebank et Multitag. L'analyse de ce corpus nous permet de caractériser le domaine clinique et de dégager les points clés pour l'adaptation d'outils d'analyse morpho-syntaxique à ce domaine. Nous montrons également les limites d'un outil entraîné sur un corpus journalistique appliqué au domaine clinique. En perspective de ce travail, nous envisageons une application du corpus clinique annoté pour améliorer l'étiquetage morpho-syntaxique des documents cliniques en français.

Abstract.
Part of Speech tagging for specialized domains : a case study with clinical documents in French.
Part-of-Speech (PoS) tagging is a core task in Natural Language Processing, often used as a stepping stone to perform more complex tasks such as information extraction or machine translation. PoS tagging of specialized documents is often challenging due to the limited availability of tools and annotated corpora dedicated to specialized domains. Herein, we present the development of a PoS annotated corpus of clinical documents in French, using annotation guidelines from the FrenchTree Bank and Multitag datasets. Through analysis of the annotated corpus, we characterize the clinical domain, including specific targets for domain adaptation. We also show the limitations of a PoS tagger trained on news documents when applied to clinical text. We expect that the domain-specific resource presented in this paper will contribute to improve PoS tagging for clinical documents in French.

Mots-clés : adaptation ; analyse morpho-syntaxique ; langue de spécialité ; dossier électronique patient.

Keywords: domain adaptation ; part-of-speech tagging ; specialized domain ; EHR.

1 Introduction

1.1 Contexte et motivation

L'étiquetage morpho-syntaxique est une tache fondamentale du Traitement Automatique de la Langue, sur laquelle reposent souvent des traitements plus complexes tels que l'extraction d'information ou la traduction automatique. Les méthodes d'apprentissage statistiques ont fortement progressé ces dernières années, mais restent limitées en domaine de spécialité par la disponibilité de corpus annotés spécifiques au domaine. Sur l'anglais, des travaux utilisant des corpus cliniques (Pakhomov *et al.*, 2006; Liu *et al.*, 2007; Ferraro *et al.*, 2013) montrent le potentiel d'adaptation à ce domaine lorsque des corpus spécialisés, même de taille modeste, sont disponibles. Une étude avec l'outil statistique MedPOST caractérise plus spécifiquement l'apport des données annotées et des ressources lexicales spécialisées riches pour l'adaption d'un étiqueteur au domaine biomédical (Smith *et al.*, 2005). A partir de ces résultats, Pécheux et al. (Pécheux *et al.*, 2014) montrent qu'un gain de performance significatif peut être obtenu dans un système de traduction automatique de documents biomédicaux utilisant un étiqueteur morpho-syntaxique adapté au domaine. Pour le français, le développement du corpus Sequoïa (Candito & Seddah, 2012) a abordé l'adaptation en domaine pour des outils d'analyse syntaxique. Le corpus comporte notamment deux rapports publics européens discutant la mise sur le marché d'un médicament. Cependant, la principale ressource annotée morpho-syntaxiquement pour le français reste le corpus journalistique French Tree Bank (Abeillé *et al.*, 2003), à la fois en terme de taille du corpus et de complexité du jeu d'étiquettes utilisé.

1.2 Objectif et contribution

Dans cet article nous nous intéressons à l'étiquetage morpho-syntaxique de textes en domaine de spécialité, et plus particulièrement les documents cliniques issus des dossiers électroniques de patients en français. A travers le développement d'un corpus clinique du français annoté morpho-syntaxiquement, nous présentons une analyse du corpus permettant de caractériser le domaine clinique et de dégager les points-clés pour l'adaptation d'outils d'analyse morpho-syntaxique à ce domaine. Le corpus annoté constitue une contribution importante de notre travail, et a vocation par la suite d'entrainer et d'évaluer un outil d'étiquetage spécialisé.

2 Méthodes

2.1 Présentation du corpus de travail

Pour cette étude, nous avons utilisé des documents d'un corpus de textes cliniques issus d'un groupe d'institutions hospitalières françaises. Ce corpus contient plusieurs types de documents (principalement des courriers, des comptes-rendus de séjour, des comptes-rendus d'actes et des ordonnances). Pour ce travail, nous avons sélectionné des documents issus du service d'hépato-gastro-nutrition faisant déjà l'objet d'une dé-identification (Grouin & Névéol, 2014) et d'autres études (Deléger & Névéol, 2014) afin d'enrichir le corpus au niveau morpho-syntaxique.

2.2 Schéma d'annotation morpho-syntaxiques

Notre choix d'étiquettes a été guidé par l'état de l'art en annotation morpho-syntaxique pour le français, exposé dans le guide d'annotation French Tree Bank ((Abeillé *et al.*, 2003) dénommé "FTB" dans la suite de cet article) et le guide d'annotation Multitag élaboré dans le cadre de la campagne PASSAGE (Villemonte De La Clergerie *et al.*, 2008).

Le schéma d'annotation du FTB est très riche et détaillé comme on peut le voir sur l'exemple suivant :

FTB :	CL-suj-1mp	V-P-1p	D-ind-ms	N-m-s
	Nous	administrons	un	traitement

Ce schéma est composé de 15 catégories lexicales et 38 sous-catégories ainsi que d'un grand nombre de traits morphologiques pour toutes les formes fléchies. Au contraire, le schéma d'annotation Multitag est principalement restreint à la syntaxe et se compose de 11 catégories lexicales et 33 sous-catégories. Il encode peu de traits morphologiques comme on peut le voir sur l'exemple suivant :

Passage :	Pp	Vm	Da	Nc
	Nous	administrons	un	traitement

La complexité du schéma du FTB demande de plus grandes compétences pour l'annotateur et augmente à la fois le temps d'annotation et le risque d'erreurs.Il semble donc plus pertinent d'utiliser ce second schéma.

Le corpus FTB a été converti au schéma Multitag ainsi que la partie médicale (notices de médicaments éditées par l'Agence Européenne du Médicament, EMEA) du corpus Sequoïa (Candito & Seddah, 2012). L'ensemble de ces corpus ont été utilisés pour entraîner un modèle CRF à l'aide de l'outil Wapiti tel que décrit dans (Lavergne *et al.*, 2010) (les caractéristiques utilisées y sont décrites à la section 5.1.2) afin de pré-annoter les documents médicaux. Le tokenizer utilisé lors de la première phase d'annotation est un outil maison suivant une segmentation proche de celui du FTB. Il n'est pas adapté au domaine médical, il a été important de modifier la segmentation manuellement afin d'obtenir une annotation morpho-syntaxique de qualité optimale.Les corpus FTB et PASSAGE sont constitués de textes journalistiques (articles du journal Le Monde) et littéraires. Le corpus EMEA est en revanche plus proche du domaine médical mais est de taille relativement réduite. Le tableau 2 présente l'ensemble du jeu d'étiquette que nous utilisons.

2.3 Pré-traitement du corpus de travail

Des travaux sur le développement de corpus clinique en anglais annoté morpho-syntaxiquement ont montré qu'il était possible de minimiser la taille du corpus annoté grâce à des heuristiques de fréquence simple (Liu *et al.*, 2007). Par ailleurs, les études précédentes menées sur notre corpus ont montré que certaines parties des documents étaient redondantes et

Etiquette complète	Etiquette réduite	Etiquette complète	Etiquette réduite
Adjectif qualificatif	Aq	Adjectif ordinal	Ao
Adjectif cardinal	Ak	Adjectif indéfini	Ai
Adjectif interrogatif	At	Adjectif possessif	Ap
Conjonction de coordination	CC	Conjonction de subordination	CS
Article	Da	Déterminant démonstratif	Dd
Déterminant indéfini	Di	Déterminant cardinal	Dk
Déterminant relatif	Dr	Déterminant possessif	Dp
Déterminant exclamatif ou interrogatif	Dt	Mot-phrase	I
Nom commun	NC	Nom propre	NP
Nom cardinal	Nk	Pronom démonstratif	Pd
Pronom indéfini	Pi	Pronom cardinal	Pk
Pronom personnel	Pp	Pronom relatif	Pr
Pronom possessif	Ps	Pronom interrogatif	Pt
Pronom réfléchi	Pf	Adverbe	Qg
Adverbe exclamatif ou interrogatif	Qx	Particule négative	Qn
Préposition	Sp	Introducteur	Sd
Verbe plein	Vm	Verbe auxiliaire	Va
Résidu	X	Ponctuations	F

TABLE 1 – Jeu d'étiquette pour l'annotation morpho-syntaxique

pouvaient présenter un intérêt limité pour l'analyse clinique et morpho-syntaxique (par exemple, en-têtes des documents). Ainsi, afin d'optimiser les efforts d'annotation dans notre étude, nous avons procédé à une sélection de phrases à l'intérieur des documents afin de concentrer le travail sur des phrases pertinentes et non-redondantes entre elles. La sélection a été opérée à l'aide d'un outil libre développé par Cohen et al. (Cohen *et al.*, 2013) en choisissant 20% comme taux de similarité maximum acceptable (valeur défaut recommandée par les auteurs) et 10 caractères comme taille des segments sur lesquels se fonde la comparaison. La sélection conduit à retenir moins de 50% de phrases ; cela s'explique en partie par le fait que les phrases "courtes" (typiquement contenues dans les en-têtes et pieds de page) de longueur inférieure à la taille des segments sont systématiquement exclues de la sélection. Les paires de phrases dont la similarité est au-dessus du seuil comprennent par exemple des variantes des phrases décrivant des techniques d'examen reprises plus ou moins à l'identique dans plusieurs documents ; ou alors des phrases comme "compte rendu d'hospitalisation de Nom Prénom, Né(e) le Date."

Un jeu de 60 documents a été préannoté avec le modèle décrit à la section précédente. Seules les phrases sélectionnées ont été pré-annotées, et présentées à un linguiste pour correction à l'aide de l'outil BRAT (Stenetorp *et al.*, 2012).

3 Annotation morphosyntaxique du corpus

3.1 Déroulement du travail d'annotation

Le travail d'annotation est fait par un seul et unique linguiste : ce choix a été imposé par le fait que nous n'avions pas d'autres linguistes disponibles à cette période. Il est prévu par la suite d'intégrer un deuxième annotateur qui viendra étayer le travail effectué en amont. Nous sommes tout à fait conscient du biais modéré que cette configuration induit, cependant il est à préciser que le linguiste ne disposait d'aucune information concernant le corpus sur lequel il a travaillé (préparation et taille des données).

Le travail d'annotation effectué sur le corpus s'est déroulé en deux temps : dans un premier temps, le linguiste a travaillé sur le corpus pré-annoté automatiquement, afin de valider ou corriger l'étiquetage proposé par l'outil statistique. Dans cette partie du travail, seule une sélection de phrases sont découpées en tokens de manière similaire au corpus FTB et pré-annotées à l'aide du modèle CRF. Les phrases sont cependant présentées à l'intérieur du document d'origine afin de fournir tous les éléments de contexte nécessaires à l'annotateur, qui peut modifier les étiquettes, sans changer le découpage en tokens proposé par l'outil. Suite à cette première phase d'annotation, une première analyse d'erreurs a permis d'identifier notamment des problèmes liés à la tokenisation du corpus.

En conséquence, dans un deuxième temps, un nouveau jeu de 30 documents vierges (sans pré-annotation) est présenté au linguiste qui peut alors effectuer l'annotation morpho-syntaxique en choisissant librement le découpage en token. Lors de cette phase, le temps moyen d'annotation a augmenté de 50% de par l'absence de pré-annotation et la nécessité de corriger les erreurs de tokenisation. Une partie des phrases annotées lors de cette étape est commune avec l'étape précédente afin de pouvoir évaluer l'impact de la pré-annotation.

Enfin, une partie des documents (10 documents) étant commune aux deux phases d'annotation, un consensus entre les deux annotations a été réalisé afin de déterminer l'étiquetage final de ces documents. Afin de limiter le biais de l'annotateur dans cette phase du travail, les documents communs ont été présentés sans distinction particulière. De plus, l'annotateur ne savait pas a priori que certains documents avaient déjà été utilisés dans la phase précédente. Interrogé sur ce point, il a indiqué ne pas s'en être rendu compte. Lors de ce consensus, les deux tokenisations et étiquetages sont présentés au linguiste qui peut choisir l'une ou l'autre pour chaque segment de phrase. L'accord intra-annotateur sur cette sous-partie du corpus est de 0.84 (précision), malgré les divergences de tokenisation.

Le tableau 2 présente la taille du corpus annoté final. A titre indicatif, nous indiquons également les chiffres équivalents pour les corpus Sequoïa-EMEA et FTB. Il est intéressant d'observer que le taux de redondance des tokens dans notre corpus est relativement faible : en moyenne, 4,31 occurrences par mot de vocabulaire contre 7,96 dans EMEA et 19,99 dans FTB. Cela atteste du succès de notre méthode de sélection de phrases non-redondantes.

	Corpus Clinique	**Sequoïa-EMEA**	**FTB**
Nombre de fichiers	80	2	44
Nombre de phrases annotées	722	1 108	11 116
Nombre de tokens annotés	13 721	22 275	679 730
Taille du vocabulaire	3 181	2 797	33 988

TABLE 2 – Statistiques descriptives du corpus clinique

3.2 Performances du système de pré-annotation

La motivation de cet étiquetage étant de caractériser les particularités du domaine médical et de fournir un corpus permettant l'adaptation des modèles existants, la première étape d'analyse a consisté à évaluer les performances du système de pré-annotation.

Le système CRF atteint 0.80 de taux d'erreur sur l'ensemble des étiquettes ce qui est relativement faible pour une tâche d'analyse morpho-syntaxique générale mais est raisonable pour un système non adapté. Pour comparaison, le système CRF atteint 0.964 de taux d'erreur sur des documents de test issus du FTB et un système entrainé uniquement sur les corpus FTB et passage obtient quant à lui 0.968 sur ce même test ce qui est comparable à l'état-de-l'art (Constant *et al.*, 2011; Denis & Sagot, 2012). Une chute de performance pouvant atteindre jusqu'à 15% est également constatée sur l'anglais, lorsque des étiqueteurs génériques sont appliqués sur des textes cliniques (Ferraro *et al.*, 2013). Une analyse plus fine des erreurs montre que la majorité des erreurs sont dues soit à des erreurs de tokenisation, soit à des particularités du domaine médical. Parmi ces dernières on retrouve notamment des confusions entre noms communs et noms propres dans les noms de procédures ou dispositifs médicaux. Par exemple, dans le syntagme *confection d'un Hartmann*, le dernier token doit être annoté comme un nom commun contrairement à ce que la majuscule suggère car il s'agit d'une marque de dispositifs médicaux (cas similaire à l'emploi du terme *Frigidaire* dans la langue générale). On retrouve aussi des mots composés néoclassiques : *thoraco-abdomino-pelvien*, des noms d'appareils : *Endoscope Olympus XQ30*) ou des noms de médicaments : *Interféron pégylé*, composés de tokens spécifiques au domaine médical.

Au final, nous calculons qu'environ 41% des erreurs de pré-annotation sont dues à des particularités de tokenisation et 22% sont liées au vocabulaire médical. En dehors de ces deux classes d'erreurs, qui sont analysées plus finement dans les sections suivantes, la pré-annotation montre des résultats de bonne qualité. Ces observations indiquent qu'avec un tokeniseur adapté, l'utilisation d'une pré-annotation permet un gain de temps significatif. Cela rejoint les résultats de travaux antérieurs sur l'annotation de corpus (Névéol *et al.*, 2011).

4 Analyse d'erreurs

4.1 Difficultés liées à la tokenisation

Parmi les phrases à étiqueter qui ont été préalablement sélectionnées, nous avons rencontré une quantité importante d'abréviations, de termes contenant des éléments de ponctuation et d'autres termes qui à première vue ressembleraient à des sigles. Ces éléments sont des formes d'expressions propres au domaine médical : ce sont des abréviations de noms de procédures médicales, des descriptions de stades d'évolution de maladies, des modalités d'administration de la prise de médicaments, et des mesures. Ce qui fait la singularité de ces termes, c'est qu'ils suivent une structure orthographique étrangère à ce qui est connu du CRF (rappelons qu'il a été exclusivement entraîné sur du corpus journalistique) ; de surcroît, la quantité de sigles est ici beaucoup plus élevée que d'ordinaire (toujours en comparaison avec le contenu d'un corpus générique).

Abréviation des termes médicaux. Les textes du domaine médical et en particulier les comptes rendus cliniques étudiés ici comportent une grande part d'abréviations. Ce phénomène, largement étudié du point de vue de la désambigüisation lexicale (Stevenson *et al.*, 2009), est dû à l'abondance de termes spécialisés ainsi qu'à la rédaction de type prise de notes.

On trouve deux types d'abréviations. D'une part des abréviations partielles, telles que *chir.* pour *chirurgie*, qui ne font pas partie des listes classiques d'abréviations et sont donc inconnues à la fois du tokeniseur et du modèle CRF. La ponctuation qui doit être considérée comme faisant partie du token est généralement séparée car reconnue comme une ponctuation finale. Le token se trouve donc incorrectement étiqueté et le marqueur de fin de phrase à tendance à propager cette erreur aux mots suivants. D'autre part, de nombreux termes médicaux tels que *anesthésie générale* ou *sérum glutamooxaloacétate transférase* sont complètement abrégés en *A.G.* et *SGOT*. Même si la morphologie de ces tokens permet de les reconnaitre plus simplement, leur regroupement sous une seule étiquette « abréviation » est ici peu approprié, ces termes étant souvent porteurs d'une information sémantique importante pour l'analyse des documents médicaux. Le choix le plus approprié est de réaliser une tokenisation assurant un découpage en tokens similaire à ceux du terme non abrégé ainsi qu'un étiquetage complet de la séquence. L'abréviation *A.G.* est donc annotée *A.* :NC *G.* :ADJq au contraire de *A.G.* :NP suggéré par le système de pré-annotation.

Mesures et abréviations complexes. Une deuxième particularité du domaine médical est l'abondance de quantités et de mesures. Si certaines sont simples, comme *3mm*, et sont correctement analysées par un système non-adapté au domaine, ce n'est pas le cas pour les plus complexes telles que *3x/j* ou *5,4 mmoles/l*. De plus, une même mesure peut-être abrégée de manière différentes, pour *3 fois par jour* par exemple, nous avons observé les abréviations suivantes dans notre corpus : *3 fois/jours*, *3 fois/j*, *3x/j*, *3/j*. . .

On trouve aussi des termes ressemblant à la fois à des abréviations et des mesures tels que *T2N+* dans *lésion du rectum T2N+* qui indique le degré d'évolution d'un cancer. Il s'agit de la classification TNM (« tumor, nodes, metastasis » en anglais) qui prend en compte la taille et la localisation de la tumeur primitive (notée parfois pT), le nombre et le site des ganglions lymphatiques régionaux qui contiennent des cellules cancéreuses, et la propagation du cancer, ou métastases, vers une autre partie du corps.

Ces deux types de termes : *3x/j* et *T2N+*, ont tendance à être considérés comme un seul token par la chaîne de traitement non-adaptée au domaine. Comme pour les abréviations simples, il est pourtant pertinent ici de les décomposer afin de les étiqueter de manière similaire à leur écriture non-abrégée.

On annotera donc *3* :Det *x* :NC /:Prep *j* : :NC et *T* :NC *2* :ADJc *N* :NC + :ADV. Cette annotation complète bien que plus coûteuse et demandant des connaissances médicales dans certains cas permet de faciliter les étapes suivantes de l'analyse automatique de ces documents.

4.2 Difficultés liées au vocabulaire

Le domaine médical a un vocabulaire très riche qui met facilement le système de pré-annotation en difficulté. On a pu noter principalement deux formes de problèmes : ceux liés à des mots spécifiques au domaine et donc inconnus du système, et ceux liés à une utilisation différente dans le domaine médical de mots classiques et donc connus.

Termes médicaux et mots hors vocabulaire. Le système de pré-annotation a été entraîné sur des corpus de textes journalistiques. Pour ce type d'étiquetage, si le corpus d'entraînement est suffisamment important, la très grande majorité des mots hors-vocabulaires seront des noms propres. On peut aussi trouver des variantes morphologiques de mots connus,

comme une forme conjuguée non vue à l'apprentissage d'un verbe connu, mais le système peut facilement gérer ces cas en exploitant le lemme du terme. La régularisation ℓ_1 utilisée à l'apprentissage du modèle CRF induit une sélection de caractéristiques. Cela a notamment pour effet de pénaliser l'utilisation des caractéristiques lexicales pour les mots les moins fréquents au profit des caractéristiques non-lexicales. Au décodage, cet effet se traduit par un fort biais du modèle vers l'étiquette `NP` pour les mots inconnus qui n'ont aucune caractéristique lexicale active.

Ces mots sont très fréquents dans les documents médicaux et les erreurs qu'ils engendrent, de par la structure du modèle CRF, se propagent facilement aux mots voisins. Pour le syntagme *microfilaments de type actine-myosine* par exemple, seul le token *de* sera correctement étiqueté car il appartient à une catégorie grammaticale fermée. Les deux tokens extrêmes sont par contre hors-vocabulaires et donc mal étiquetés et ces erreurs se propagent sur le token *type* pourtant connu du système.

Il est à noter que ces mots hors-vocabulaires ne peuvent être traités uniquement grâce à l'ajout de données médicales étiquetées. Il est en effet impossible d'en avoir une couverture suffisante et l'utilisation de lexiques se trouve ici indispensable.

Homographies. L'homographie de certains termes existant en dehors du domaine médical a aussi été un facteur d'erreurs d'étiquetage. De nombreux termes sont utilisés différemment en contexte médical et leur catégorie grammaticale change, par exemple dans le syntagme *veine porte*, le terme *porte* est un adjectif qualificatif, alors qu'il est un nom commun ou une forme conjuguée du verbe *porter* dans les textes du corpus d'entrainement.

De même, le terme *scanner* est un nom commun dans le domaine médical. S'il peut aussi être un nom commun dans le domaine général, il est plus fréquemment étiqueté comme verbe, ce qui conduit à des erreurs. Une étude sur notre corpus montre que parmi ces homographies, dans 23% des cas, la catégorie grammaticale utilisée en domaine médicale est inobservée dans le corpus général et donc impossible à prédire pour le système.

Concernant l'homographie, nous avons aussi pu observer des erreurs plus complexes. Par exemple, le syntagme *le toucher rectal* peut être étiqueté `PROp VRB ADV` ou `DET NC ADJ`. En dehors de tout contexte, les deux choix d'annotations sont plausibles, mais dans le vocabulaire médical, *toucher* désigne une pratique. Ce terme est utilisé en tant que nom commun et seule la deuxième possibilité est donc acceptable.

Ambiguïté des noms de médicaments. Les noms de médicaments tels que *Dafalgan effervescent*, *Fudicine pommade* et *Néoral 50* présentent une difficulté supplémentaire. Deux étiquetages sont possibles pour chacun d'eux : soit les deux tokens forment un nom propre et reçoivent tous les deux le tag `NP`, soit seul le premier token est considéré comme constituant le nom du médicament, le deuxième étant un adjectif qui en précise la forme. Dans ce deuxième cas, le deuxième token recevra le tag `ADJq` ou `ADJc` suivant les cas.

Ces deux schémas d'annotation sont justifiables mais il est nécessaire d'assurer la cohérence du choix sur tout le corpus. Le système de pré-annotation est incohérent ici car, lorsque l'adjectif ne fait pas partie de son vocabulaire il choisit la première forme, mais dans le cas contraire ou pour les numéraux, c'est la deuxième forme qui l'emporte.

Médicalement, le deuxième token correspond à une caratéristique associée au médicament, telle que la concentration, la forme ou la voie d'administration. Ainsi, il semble plus pertinent de ce point de vue de choisir la deuxième solution d'étiquetage.

5 Conclusion et perspectives

L'analyse d'un corpus clinique du français annoté morpho-syntaxiquement et en particulier des erreurs d'étiquetage faites par un outil générique a mis en évidence deux principales caractéristiques du domaine clinique. D'une part, un *vocabulaire spécialisé* qui dénote des connaissances à apporter à l'étiqueteur grâce à des lexiques spécialisés et des données étiquetées. D'autre part, un besoin d'une *tokénisation particulière* pour des phénomènes linguistiques particulièrement prévalents dans les textes cliniques, tels que les posologies, mesures et abréviations. Pour permettre à un outil statistique d'apprendre ce type de construction, il est nécessaire de disposer d'une quantité satisfaisante de données annotées. Une contribution importante du travail présenté dans cet article est le développement d'un corpus du domaine clinique annoté morpho-syntaxiquement, dans le but d'entrainer et d'évaluer un outil d'étiquetage spécialisé. Le développement d'un tel outil est en cours. Une perspective à moyen terme est d'enrichir le corpus existant avec l'intervention d'un deuxième annotateur linguiste afin de faire de ce corpus une référence de qualité destinée à évaluer différents outils. Cet objectif reste conditionné à l'obtention d'une autorisation de la CNIL en raison de la nature sensible des documents.

Remerciements

Nous remercions le Service d'Informatique Biomédicale (SIBM) ainsi que l'équipe CISMeF du CHU de Rouen qui nous ont permis d'utiliser le corpus LERUDI pour cette étude. Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence CABeRneT ANR-13-JS02-0009-01.

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for French. In A. ABEILLÉ, Ed., *Treebanks*. Kluwer, Dordrecht.

CANDITO M.-H. & SEDDAH D. (2012). Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Actes de TALN 2012*, p. 321–334.

COHEN R., ELHADAD M. & ELHADAD N. (2013). Redundancy in electronic health record corpora : analysis, impact on text mining performance and mitigation strategies. *BMC Bioinformatics*, **14**, 10.

CONSTANT M., TELLIER I., DUCHIER D., DUPONT Y., SIGOGNE A. & BILLOT S. (2011). Intégrer des connaissances linguistiques dans un crf : application à l'apprentissage d'un segmenteur-étiqueteur du français. In *Actes de TALN 2011*.

DELÉGER L. & NÉVÉOL A. (2014). Identification automatique de zones dans des documents pour la constitution d'un corpus médical en français. In *Actes de TALN 2014*, p. 568–573.

DENIS P. & SAGOT B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art pos tagging. *Lang. Resour. Eval.*, **46**(4), 721–736.

FERRARO J., DAUMÉ H., DUVALL S., CHAPMAN W., HARKEMA H. & HAUG P. (2013). Improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation. *J Am Med Inform Assoc*, **20**(5), 931–939.

GROUIN C. & NÉVÉOL A. (2014). De-identification of clinical notes in french : towards a protocol for reference corpus developpement. *J Biomed Inform*, **50**, 151–61.

LAVERGNE T., CAPPÉ O. & YVON F. C. (2010). Practical very large scale crfs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 504–513, Uppsala, Sweden.

LIU K., CHAPMAN W., HWA R. & CROWLEY R. (2007). Heuristic sample selection to minimize reference standard training set for a part-of-speech tagger. *J Am Med Inform Assoc*, **14**(5), 641–650.

NÉVÉOL A., DOĞAN R. I. & LU Z. (2011). Semi-automatic semantic annotation of PubMed queries : a study on quality, efficiency, satisfaction. *J Biomed Inform*, **44**(2), 310–8.

PAKHOMOV S., CODEN A. & CHUTE C. (2006). Developing a corpus of clinical notes manually annotated for part-of-speech. *International Journal of Medical Informatics*, **75**(6), 418–429.

PÉCHEUX N., GONG L., DO Q. K., MARIE B., IVANISHCHEVA Y., ALLAUZEN A., LAVERGNE T., NIEHUES J., MAX A. & YVON F. (2014). Limsi @ wmt'14 medical translation task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, p. 246–253, Baltimore, Maryland, USA : Association for Computational Linguistics.

SMITH L., RINDFLESCH T. & WILBUR W. (2005). The importance of the lexicon in tagging biological text. *Natural Language Engineering*, **12**(2), 1–17.

STENETORP P., PYYSALO S., TOPIĆ G., OHTA T., ANANIADOU S. & TSUJII J. (2012). Brat : A web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, p. 102–107, Stroudsburg, PA, USA : Association for Computational Linguistics.

STEVENSON M., GUO Y., AL AMRI A. & GAIZAUSKAS R. (2009). Disambiguation of biomedical abbreviations. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '09, p. 71–79, Stroudsburg, PA, USA : Association for Computational Linguistics.

VILLEMONTE DE LA CLERGERIE E., HAMON O., MOSTEFA D., AYACHE C., PAROUBEK P. & VILNAT A. (2008). Passage : from french parser evaluation to large sized treebank. In *Proceedings of the 6th International Conference on Languages Resources and Evaluation, LREC 2008*, Marrakech, Morocco.