Étiquetage multilingue en parties du discours avec MElt

Benoît Sagot

Étiquetage multilingue en parties du discours avec MElt

Benoît Sagot
ALPAGE, Inria & Université Paris-Diderot, Paris
benoit.sagot@inria.fr

RÉSUMÉ

Nous présentons des travaux récents réalisés autour de MElt, système discriminant d'étiquetage en parties du discours. MElt met l'accent sur l'exploitation optimale d'informations lexicales externes pour améliorer les performances des étiqueteurs par rapport aux modèles entraînés seulement sur des corpus annotés. Nous avons entraîné MElt sur plus d'une quarantaine de jeux de données couvrant plus d'une trentaine de langues. Comparé au système état-de-l'art MarMoT, MElt obtient en moyenne des résultats légèrement moins bons en l'absence de lexique externe, mais meilleurs lorsque de telles ressources sont disponibles, produisant ainsi des étiqueteurs état-de-l'art pour plusieurs langues.

ABSTRACT

Multilingual part-of-speech tagging with MElt.

We describe recent evolutions of MElt, a discriminative part-of-speech tagging system. MElt is targeted at the optimal exploitation of information provided by external lexicons for improving its performance over models trained solely on annotated corpora. We have trained MElt on more than 40 datasets covering over 30 languages. Compared with the state-of-the-art system MarMoT, MElt's results are slightly worse on average when no external lexicon is used, but slightly better when such resources are available, resulting in state-of-the-art taggers for a number of languages.

MOTS-CLÉS : Étiquetage morphosyntaxique, Multilingue, Évaluation.

KEYWORDS: Part-of-speech tagging, Multilingual, Evaluation.

1 Introduction

L'étiquetage morphosyntaxique (*tagging*) est une tâche désormais classique en traitement automatique des langues, pour laquelle de nombreux systèmes ont été développés ou adaptés à un large éventail de langues. Elle consiste à associer à chaque « mot » une *étiquette morphosyntaxique* dont la granularité peut aller d'une simple catégorie morphosyntaxique, ou partie du discours, à une catégorie plus fine et enrichie par des traits morphologiques (genre, nombre, cas, temps, mode, etc.).

L'utilisation d'algorithmes d'apprentissage automatique faisant usage de corpus annotés manuellement est aujourd'hui la norme pour le développement d'étiqueteurs morphologiques. Différents types d'algorithmes ont été utilisés, dont successivement les modèles de Markov cachés bigrammes puis trigrammes (Merialdo, 1994; Brants, 1996, 2000), les arbres de décision (Schmid, 1994; Magerman, 1995), les modèles de Markov à maximisation d'entropie (Maximum Entropy Markov Models, MEMM ; Ratnaparkhi, 1996) et les champs aléatoires conditionnels (Conditional Random Fields, CRF ; Constant *et al.*, 2011). Ces algorithmes d'apprentissage automatique permettent la construction de systèmes d'étiquetage pour n'importe quelle langue, pourvu que l'on dispose de

données d'apprentissage adaptées. Certains systèmes sont largement utilisés, précisément parce qu'ils ont été utilisés pour entraîner de nombreux modèles faciles à télécharger et à utiliser. C'est par exemple le cas de TreeTagger [1] (Schmid, 1994), qui repose sur l'apprentissage d'arbres de décision, pour lequel des modèles pour plus d'une vingtaine de langues sont disponibles. Le modèle fourni pour le français constitue d'ailleurs probablement l'étiqueteur le plus utilisé dans la communauté francophone, certainement en raison de sa gratuité [2], de son ancienneté et de sa facilité d'utilisation, malgré des performances inférieures à plusieurs étiqueteurs plus récents (Denis & Sagot, 2009).

Pour compléter les données annotées, plusieurs travaux ont montré la pertinence de s'appuyer sur des informations lexicales externes, et notamment sur des lexiques morphosyntaxiques associant par exemple une partie du discours à un grand nombre de mots. Ces informations peuvent être utilisées sous forme de contraintes au moment de l'étiquetage de nouvelles données (Kim *et al.*, 1999; Hajič, 2000) ou pour l'extraction de traits qui complètent les traits extraits du corpus annoté dès la phase l'apprentissage (cf. par exemple Chrupała *et al.*, 2008). C'est cette dernière approche qui a été explorée par les développeurs du système d'étiquetage MElt (Denis & Sagot, 2009, 2010, 2012), système dont les versions précédentes étaient état-de-l'art sur le français — plus précisément, sur la tâche d'étiquetage correspondant à la version dite FTB-UC du French TreeBank (cf. ci-dessous). Toutefois, ce choix d'évaluation est discutable, comme nous le verrons à la section 2. Nous rappelons à la section 3 les grandes lignes du fonctionnement de MElt et nous décrirons brièvement les améliorations qui lui ont été récemment apportées. Nous présentons ensuite à la section 4 les nombreux modèles d'étiquetage que nous avons entraînés pour une quarantaine de langues dont naturellement le français, notamment grâce aux corpus du projet *Universal Dependencies* et à ceux développés dans le cadre des campagnes SPMRL (Seddah *et al.*, 2013), souvent secondées par un lexique externe. Nous discutons des performances de ces modèles, y compris par leur comparaison avec ceux produits par le système MarMoT (Müller *et al.*, 2013; Müller & Schütze, 2015), probablement le plus performant actuellement, à partir des mêmes données.

2 Évaluation des étiqueteurs morphosyntaxiques

Évaluer un étiqueteur morphosyntaxique consiste simplement à mesurer son taux d'exactitude, c'est-à-dire le pourcentage de « mots » ayant reçu la bonne étiquette. Ce pourcentage est souvent complété par l'exactitude sur les seuls mots inconnus (absents du corpus d'apprentissage). Toutefois, il faut garder à l'esprit que les comparaisons que l'ont peut faire de cette façon entre différents systèmes n'évaluent pas vraiment les systèmes en soi — quoi que cela puisse signifier —, mais plutôt leur adéquation au corpus utilisé, avec toutes ses caractéristiques et notamment sa taille, son homogénéité ou encore les caractéristiques de son jeu d'étiquettes (nombre d'étiquettes distinctes...).

Par ailleurs, un étiqueteur morphosyntaxique est destiné à être utilisé *in fine* sur des corpus bruts, ce qui nécessite un découpage préalable en unités élémentaires destinées à recevoir des étiquettes. On peut considérer que de telles unités élémentaires, en tant qu'elles seront associées à des étiquettes linguistiquement signifiantes, doivent être des unités linguistiquement valides et donc correspondre à des unités lexicales [3]. Mais la détection des composés est une tâche difficile. Il est donc bien plus simple et bien plus fréquent de ramener la tâche de découpage en unités élémentaire à une tâche de découpage en tokens. Dans ce cas, associer des étiquettes linguistiques à des tokens est nécessairement

1. Disponible sur `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`.
2. TreeTagger est un outil gratuit mais pas libre, puisque son code source n'est pas disponible.
3. Dans ce cas, les composés sont considérés comme une unité unique, et les amalgames comme reflétant plusieurs unités.

une approximation dès lors que la correspondance entre tokens et unités lexicales n'est pas biunivoque. Pour contourner en partie cette difficulté, on s'appuie généralement sur le fait que les composés ont souvent des composants identifiables comme relevant d'une catégorie connue [4]. Naturellement, une telle approche nécessite que les corpus annotés sur lesquels les modèles sont entraînés et évalués reflètent ce même choix.

Ce n'est pourtant pas la stratégie utilisée par la majorité des travaux antérieurs sur l'étiquetage morphosyntaxique du français. En effet, si le corpus de référence pour le français est le French TreeBank (FTB ; Abeillé *et al.*, 2003), il est le plus souvent utilisé dans les travaux sur l'étiquetage morphosyntaxique dans sa version dite FTB-UC (Candito & Crabbé, 2009). Cette version diffère du FTB originel en ceci que tous les composés qui ne correspondent pas à une séquence régulière de parties du discours voient leurs tokens constitutifs fusionnés en une unité élémentaire unique (pour les autres composés, les unités sont les tokens). Il s'agit donc d'une segmentation intermédiaire entre les deux approches décrites ci-dessus. À l'inverse, la variante FTB-SPMRL, développée dans le cadre de la campagne SPMRL d'évaluation des analyseurs syntaxiques pour des langues à morphologie riche (Seddah *et al.*, 2013), utilise systématiquement les tokens comme unités élémentaires [5]. De plus le FTB-SPMRL contient des données supplémentaires : il dispose de 5 000 phrases de plus pour l'entraînement et ses sections de développement et d'évaluation sont environ deux fois plus importantes que dans le FTB-UC. C'est donc l'évaluation sur le FTB-SPMRL qui est la plus pertinente, et non l'évaluation sur le FTB-UC, qui a pourtant servi de référence à de nombreux travaux, y compris ceux concernant MElt [6].

3 Le système d'étiquetage MElt

MElt est un système d'étiquetage de séquences qui repose sur les modèles markoviens à maximisation d'entropie (MEMM, Ratnaparkhi, 1996), une classe de modèles discriminants adaptés aux problèmes séquentiels. Dans (Denis & Sagot, 2009, 2012), les développeurs de MElt ont étudié de quelle façon un tel modèle peut tirer parti non seulement des informations extraites du corpus d'entraînement mais également d'informations issues d'un lexique exogène à large couverture. Ils ont montré que les performances obtenues sont meilleures si l'intégration de ces dernières se fait sous forme de traits supplémentaires, par opposition à l'utilisation du lexique externe comme source de contraintes au moment de l'étiquetage (par exemple, en n'autorisant pas l'étiquetage d'un mot connu du lexique par une étiquette que le lexique ne lui associe pas). Ces traits supplémentaires, et notamment la ou les catégories fournies par le lexique pour les mots du contexte, permettent notamment une meilleure modélisation du contexte droit du mot courant, en l'absence d'étiquettes déjà attribuées par le modèle qui étiquette de gauche à droite. De plus, la gestion des mots inconnus est améliorée, un mot inconnu du corpus pouvant être connu du lexique externe.

Les expériences menées en couplant ainsi le FTB-UC et le lexique Le*fff* (Sagot, 2010) via des traits lexicaux ont conduit à un étiqueteur de niveau état-de-l'art distribué librement, dont l'exactitude était alors de 97,75% (91,36% sur les inconnus).

4. Ainsi, *pomme de terre* n'est pas difficile à analyser comme une séquence nom+préposition+nom, bien qu'il s'agisse en réalité d'un nom composé.

5. Les composés sont indiqués via des traits dont la prédiction ne fait pas partie de la tâche d'étiquetage morphosyntaxique.

6. Le FTB-UC et le FTB-SPMRL utilisent malgré tout un même jeu de 29 étiquettes qui diffère de celui du FTB originel, mis à part 4 étiquettes rares présentes uniquement dans le FTB-SPMRL.

Par rapport à la version décrite dans (Denis & Sagot, 2012), nous avons fait évoluer MElt dans trois directions principales. Tout d'abord, nous avons rajouté de nouveaux traits (préfixes et suffixes du mot à droite du mot courant) et fait varier certaines propriétés de l'espace des traits utilisés au cours d'expériences menées sur les sections d'entraînement et de développement du FTB-SPMRL. Nous avons notamment fait varier la longueur maximale des préfixes et des suffixes pris en compte par le modèle (pour le mot courant et celui à sa droite). Nous avons ensuite choisi comme nouveaux paramètres par défaut dans MElt des paramètres qui donnaient de bons résultats sur le sous-corpus de développement de FTB-SPMRL. Les expériences ci-dessous utilisent toutes ce nouveau jeu de traits.

Par ailleurs, nous avons encapsulé le moteur d'étiquetage dans des outils permettant l'identification de certains types d'entités nommées (dates, adresses…) qui sont alors considérés comme des mots uniques par le modèle d'étiquetage et dont l'étiquetage interne est effectué grâce à des règles dédiées [7].

Enfin, nous avons encapsulé le moteur d'étiquetage dans des outils permettant le traitement de corpus bruités tels que trouvés sur le web. Dans les travaux décrits ici, cette option n'a pas été activée. Pour plus de détails, on pourra se reporter par exemple à (Chanier *et al.*, 2014).

4 Expériences

Pour le français, plusieurs autres étiqueteurs morphosyntaxiques dédiés sont désormais disponibles, au-delà des systèmes d'étiquetage génériques qu'il est possible d'entraîner sur des données françaises annotées, éventuellement secondés par un lexique externe comme le L*efff*. Comme expliqué plus haut, et contrairement à la majorité des travaux antérieurs, nous avons décidé, pour le français, de comparer MElt avec d'autres étiqueteurs sur le FTB-SPMRL. À cet effet nous avons retenu TreeTagger, mentionné précédemment, et MarMoT, système reposant sur les CRF et qui peut être considéré comme le meilleur système d'étiquetage actuellement disponible (Müller *et al.*, 2013; Müller & Schütze, 2015) [8]. Les résultats de ces expériences sont fournis au tableau 1. Nous avons également encapsulé MarMoT dans nos outils de gestion de certaines entités nommées pour comparer les gains obtenus par ce biais entre MElt et MarMoT (colonnes « +EN »).

Nous avons également souhaité évaluer la portabilité de MElt à d'autres langues, et notamment étudier la systématicité de l'amélioration induite par l'utilisation d'un lexique externe. Nous avons donc cherché des ensembles de corpus pour de nombreuses langues facilement disponibles et exploitables. Or deux ensembles de corpus annotés syntaxiquement ont été développés récemment, en grande partie mais pas seulement à partir de corpus pré-existants : les corpus SPMRL [9] (Seddah *et al.*, 2013),

7. L'idée sous-jacente est que les principes linguistiques qui sous-tendent les guides d'annotation perdent de leur validité lorsque l'on est à l'intérieur d'une entité nommée, dont l'organisation est régie par des conventions spécifiques (sur quel fondement décider de la partie du discours de *2016* dans *lundi 2 janvier 2016* ?). Cela est visible dans les corpus, notamment dans le FTB, où les annotations à l'intérieur des entités nommées ne sont pas très cohérentes. Un modèle statistique tel que MElt, sans outil dédié pour la gestion des entités nommées, ne peut apprendre comment annoter de telles entités à partir d'annotations de référence incohérentes. De plus, MElt ne peut pas apprendre directement les généralisations pertinentes concernant le contexte d'apparition de tel ou tel type d'entité. À l'inverse, réunitariser certains types d'entités nommées (par exemple, remplacer toutes les dates par `_DATE`) et étiqueter les tokens qui les composent grâce à des règles systématiques permet de résoudre ces difficultés.

8. Nous aurions pu également réaliser des comparaisons avec d'autres étiqueteurs libres, spécifiques ou non au français, tels que LIA_tagg (Nasr *et al.*, 2004), le Stanford Tagger (Toutanova & Manning, 2000; Manning, 2011), LGtagger (Constant *et al.*, 2011; Constant & Sigogne, 2011) ou Morfette (Chrupała *et al.*, 2008). Pour plusieurs de ces étiqueteurs, un modèle d'étiquetage pour le français peut être téléchargé. Il s'appuie souvent en tout ou partie sur les informations lexicales du L*efff*.

9. `http://www.spmrl.org`.

dont le FTB-SPMRL fait partie, et les corpus du projet *Universal Dependencies* [10] (désormais UD), actuellement dans leur version 1.2. De chacun de ces corpus peuvent immédiatement être dérivés des corpus annotés morphosyntaxiquement. Nous avons donc mené nos expériences sur ces corpus.

Les données SPMRL couvrent 9 langues : arabe, basque, français (FTB-SPMRL), allemand, hébreu, hongrois, coréen, polonais et suédois. De son côté, le projet UD donne accès aujourd'hui à 37 corpus annotés syntaxiquement couvrant 33 langues distinctes. Contrairement aux données de la campagne SPRML, le projet UD cherche à harmoniser les annotations d'une langue à l'autre. Au niveau morphosyntaxique, cela se traduit par l'utilisation d'un jeu d'étiquette commun, les 17 *universal POS tags* [11]. La pertinence linguistique d'un tel jeu d'étiquettes conçu pour être appliqué à toutes les langues est certainement discutable, mais l'initiative UD n'en est pas moins utile en pratique.

Nous avons donc entraîné MElt et MarMoT sur tous les corpus SPMRL et UD (version 1.2) [12], ainsi que sur deux corpus de référence qui ne font pas partie des deux ensemble précédents : le Penn TreeBank pour l'anglais et le Prague Dependency Treebank (version 3.0) pour le tchèque.

Comme pour le français, nous avons fourni à MElt et à MarMoT, dès lors que cela était possible, un lexique externe comme source d'informations complémentaires. Faute de place, nous ne détaillerons pas ici l'origine de chacun de ces lexiques. Un certain nombre d'entre eux, comme le Le*fff*, ont été développés dans le formalisme Alexina (Sagot, 2010). D'autres, librement disponibles, ont été téléchargés et convertis automatiquement dans ce même formalisme, avant extraction de lexiques morphosyntaxiques (couples forme–catégorie) exploitables par MElt et MarMoT.

5 Discussion

Les résultats présentés aux tables 1 et 2 montrent plusieurs choses. Tout d'abord, MElt est, en moyenne, moins performant que MarMoT lorsqu'ils sont entraînés uniquement sur les corpus d'entraînement, sans lexique externe. Une légère corrélation peut être observée entre d'une part l'écart entre MElt et MarMoT et d'autre part la taille du corpus d'entraînement, MElt ayant tendance à être meilleur sur de petits corpus d'entraînement. Une fois prises en compte les informations lexicales externes, la situation s'inverse. L'amélioration est significative pour les deux systèmes par rapport aux modèles n'en faisant pas usage, mais cette amélioration est bien plus élevée pour MElt, au point que ce dernier passe en moyenne devant MarMoT.

Il est délicat de tirer des conclusions générales à partir de tels résultats. Il nous semble toutefois que la situation ne doit pas être très différente de l'affirmation suivante : le modèle statistique de type CRF sous-jacent à MarMoT est plus performant que le MEMM sur lequel s'appuie MElt ; toutefois, la façon dont les informations lexicales externes peuvent être intégrées aux modèles MElt est suffisamment performante pour permettre souvent à ce dernier de passer devant MarMoT en présence de telles ressources. Dans ces cas-là, le modèle produit par MElt est état de l'art.

Par ailleurs, pour le français, nos outils de gestion des entités nommées améliorent les résultats.

L'ensemble des modèles MElt ainsi entraînés est librement disponible, tout comme MElt lui-même.

10. `http://universaldependencies.org`.

11. Cf. `http://universaldependencies.org/u/pos/all.html`, raffinement de (Petrov *et al.*, 2012).

12. À quelques exceptions près. En effet, les traits utilisés par ces systèmes s'appuient en partie sur la granularité fines des systèmes d'écriture alphabétiques, qui permet d'extraire des préfixes et des suffixes pertinents. Nous n'avons donc pas pris en compte dans nos expériences les données coréennes et japonaises, dont les systèmes d'écriture ne sont pas alphabétiques.

Système	Sans lexique externe				Avec lexique externe (L*efff*)			
	standard		+EN		standard		+EN	
	Total	**Inc.**	**Total**	**Inc.**	**Total**	**Inc.**	**Total**	**Inc.**
TreeTagger	95,54	84,47			96,11	86,35		
MarMoT	97,29	86,95	97,41	87,79	**97,41**	88,08	97,50	88,41
MElt	97,12	85,54	97,16	87,80	97,36	**88,16**	**97,51**	**90,63**

TABLE 1 – Évaluation comparative de l'exactitude de MElt, de MarMoT et de TreeTagger (en %) sur le corpus FTB-SPMRL (33 étiquettes distinctes, 443 113 mots dans le corpus d'entraînement).

Langue	Corpus source	nb. mots d'entr.	nb. d'ét.	Lexique	MElt		MElt vs. MarMoT	
					Sans lex.	Av. lex.	Sans lex.	Av. lex.
Corpus du projet Universal Dependencies (version 1.2)								
Allemand (web)	UDT	274 345	16	DeLex	92,74	93,43	–0,11	**+0,33***
Anglais (web)	Engl. Web TB	204 586	17	EnLex	94,06	94,60	–0,31*	**+0,05**
Arabe	PADT	225 853	16	—	98,39	—	–0,29*	—
Basque	BDT (extr.)	72 974	16	—	94,72	—	**+0,05**	—
Bulgare	BulTreeBank	124 474	16	Multext–east	97,75	98,15	**+0,11**	**+0,10**
Croate	SEThr	78 817	14	HML	95,08	96,70	–0,07	**+0,51***
Danois	DDT	88 979	17	STO	95,48	96,30	–0,08	**+0,14**
Espagnol (web)	UDT	389 703	17	L*eff*e	95,32	95,57	**+0,18**	**+0,33**
Estonien	Arborest	7 687	15	Multext–east	89,64	94,46	**+0,52**	**0,00**
Finlandais	FinnTreeBank	127 980	15	—	93,29	—	–1,24*	—
Finlandais	Turku Dep. TB	162 721	15	—	93,24	—	–2,10*	—
Français (web)	UDT	366 138	18	L*efff*	95,81	96,14	–0,32	–0,20
Gothique	PROIEL	44 722	13	—	95,48	—	–0,22	—
Grec class.	AGDT 2.0 (extr.)	196 083	13	Diogenes	93,64	94,03	–0,34*	–0,29*
Grec class.	PROIEL	166 061	13	Diogenes	96,74	97,21	–0,33*	–0,01
Grec mod.	GrDepTB	47 449	11	DELA_gr	97,65	98,08	**0,00**	**+0,09**
Hébreu	HebConstTB	167 176	17	—	95,85	—	**+0,05**	—
Hindi	HDTB	281 057	16	—	96,29	—	–0,03	—
Hongrois	SzTB	20 764	16	Multext–east	94,42	94,86	**+0,07**	–0,04
Indonésien (web)	UDT	97 531	16	Kateglo	93,74	93,83	**+0,11**	**+0,01**
Irlandais	IDT	16 701	16	inmdb	92,38	92,75	**+0,99***	**+1,15***
Italien	(various)	265 992	18	Morph_it	97,44	97,82	–0,35*	–0,21*
Latin	Index Thom. TB	246 573	14	Diogenes	98,84	99,04	–0,16	**+0,02**
Latin	LDT	37 819	12	Diogenes	93,61	94,70	**+0,63***	**+1,16***
Latin	PROIEL	132 376	13	Diogenes	96,68	96,83	–0,07	–0,16
Néerlandais	Alpino	188 882	16	Alpino_lex	90,17	90,51	**+0,59**	**+0,36**
Norvégien	NDT	244 776	17	OrdBank	96,68	97,58	–0,58*	–0,04
Persan	UDT	122 093	16	PerLex	96,72	97,17	**+0,29***	**+0,20**
Polonais	Składnica	69 499	13	PolLex	96,12	97,77	–0,09	**+0,30**
Portugais	Flor. Sint. (extr.)	201 845	17	Labellex_pt	97,38	97,56	–0,05	**+0,17**
Roumain	RoRefTrees	9 291	17	Multext–east	91,09	94,35	**+2,02***	**+2,03***
Slovène	ssj500k	112 334	16	SloLeks	96,05	97,53	–0,18	**+0,30***
Suédois	Talbanken (extr.)	66 645	15	Saldo	95,97	96,90	–0,06	**+0,10**
Tamoul	TamilTB	6 329	14	—	89,14	—	**+1,51***	—
Tchèque	PDT	1 175 374	18	Morfflex (extr.)	98,01	98,58	–0,32*	**+0,10***
Vieux–slave	PROIEL	46 025	13	—	96,24	—	–0,02	—
Corpus de la campagne SPMRL								
Allemand	TIGER	77 220	54	DeLex	96,93	97,19	–0,43*	–0,34*
Basque	BDT	25 136	46	—	95,77	—	–0,10	—
Hébreu	HebConstTB	15 971	50	—	93,79	—	–0,30*	—
Hongrois	SzTB	40 782	23	Multext–east	96,68	96,94	–0,30*	**+0,01**
Polonais	Składnica	21 793	29	PolLex	96,44	97,55	**+0,02**	**+0,35***
Suédois	Talbanken	76 332	25	Saldo	96,60	97,54	**+0,23**	**+0,42***
Anglais	PTB	43 210	45	EnLex	96,81	97,01	–0,43*	–0,30*
Tchèque	PDT3.0	364 636	59	Morfflex (extr.)	98,65	99,29	–0,04	**+0,27***

TABLE 2 – Exactitude de MElt et de MarMoT (en %) sur divers corpus. Les deux dernières colonnes donnent l'écart entre MElt et de MarMoT. Un écart positif indique que MElt est meilleur, et est mis en évidence typographiquement ; un écart significatif ($p < 0{,}05$) est suivi d'un astérisque.

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for French. In A. ABEILLÉ, Ed., *Treebanks*. Dordrecht, Pays-Bas : Kluwer Academic Publishers.

BRANTS T. (1996). Estimating markov model structures. In *Proceedings of the Fourth Conference on Spoken Language Processing (ICSLP-96*, p. 893–896.

BRANTS T. (2000). TnT : A Statistical Part-of-speech Tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, ANLC '00, p. 224–231, Seattle, Washington, États-Unis.

CANDITO M. & CRABBÉ B. (2009). Improving generative statistical parsing with semi-supervised word clustering. In *Proceedings of the 11th International Conference on Parsing Technologies*, IWPT'09, p. 138–141, Paris, France.

CHANIER T., POUDAT C., SAGOT B., ANTONIADIS G., WIGHAM C. R., HRIBA L., LONGHI J. & SEDDAH D. (2014). The CoMeRe corpus for French : structuring and annotating heterogeneous CMC genres. *JLCL - Journal for Language Technology and Computational Linguistics*, **29**(2), 1–30.

CHRUPAŁA G., DINU G. & VAN GENABITH J. (2008). Learning morphology with morfette. In *Proceedings of the 6th Language Resource and Evaluation Conference*, Marrakech, Maroc.

CONSTANT M. & SIGOGNE A. (2011). MWU-aware part-of-speech tagging with a CRF model and lexical resources. In *Proceedings of the Workshop on Multiword Expressions : From Parsing and Generation to the Real World*, MWE '11, p. 49–56, Portland, Oregon, États-Unis.

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*, TALN 2011, p. 321–332, Montpellier, France.

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

DENIS P. & SAGOT B. (2010). Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morphosyntaxique état-de-l'art du français. In *Actes de Traitement Automatique des Langues Naturelles : TALN 2010*, Montréal, Québec, Canada.

DENIS P. & SAGOT B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. *Language Resources and Evaluation*, **46**(4), 721–736.

HAJIČ J. (2000). Morphological Tagging : Data vs. Dictionaries. In *Proceedings of ANLP'00*, p. 94–101, Seattle, Washington, États-Unis.

KIM J.-D., LEE S.-Z. & RIM H.-C. (1999). HMM Specialization with Selective Lexicalization. In *Proceedings of the join SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, EMNLP-VLC-99.

MAGERMAN D. M. (1995). Statistical decision-tree models for parsing. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, ACL '95, p. 276–283, Cambridge, Massachusetts, États-Unis.

MANNING C. D. (2011). Part-of-speech tagging from 97% to 100% : Is it time for some linguistics ? In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing'11, p. 171–189, Tokyo, Japon.

MERIALDO B. (1994). Tagging english text with a probabilistic model. *Computational Linguistics*, **20**(2), 155–171.

MÜLLER T., SCHMID H. & SCHÜTZE H. (2013). Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, p. 322–332, Seattle, Washington, États-Unis : Association for Computational Linguistics.

MÜLLER T. & SCHÜTZE H. (2015). Robust morphological tagging with word representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, Denver, Colorado, États-Unis.

NASR A., BÉCHET F. & VOLANSCHI A. (2004). Tagging with Hidden Markov Models using ambiguous tags. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Genève, Suisse.

PETROV S., DAS D. & MCDONALD R. (2012). A universal part-of-speech tagset. In N. C. C. CHAIR), K. CHOUKRI, T. DECLERCK, M. U. DOĞAN, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Eds., *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turquie.

RATNAPARKHI A. (1996). A maximum entropy model for part-of-speech tagging. In *Proceedings of International Conference on Empirical Methods in Natural Language Processing*, p. 133–142.

SAGOT B. (2010). The Le*fff*, a freely available, accurate and large-coverage lexicon for french. In *Proceedings of the 7th international conference on Language Resources and Evaluation (LREC'2010)*, La Valette, Malte.

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

SEDDAH D., TSARFATY R., KÜBLER S., CANDITO M., CHOI J. D., FARKAS R., FOSTER J., GOENAGA I., GOJENOLA GALLETEBEITIA K., GOLDBERG Y., GREEN S., HABASH N., KUHLMANN M., MAIER W., NIVRE J., PRZEPIÓRKOWSKI A., ROTH R., SEEKER W., VERSLEY Y., VINCZE V., WOLIŃSKI M., WRÓBLEWSKA A. & VILLEMONTE DE LA CLERGERIE E. (2013). Overview of the SPMRL 2013 shared task : A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, p. 146–182, Seattle, Washington, États-Unis.

TOUTANOVA K. & MANNING C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of International Conference on New Methods in Language Processing*, p. 63–70, Hong Kong, Chine.