Divergences entre annotations dans le projet *Universal Dependencies* et leur impact sur l'évaluation de l'étiquetage morpho-syntaxique

Guillaume Wisniewski François Yvon
LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91 405 Orsay, France
prénom.nom@limsi.fr

RÉSUMÉ

Ce travail montre que la dégradation des performances souvent observée lors de l'application d'un analyseur morpho-syntaxique à des données hors domaine résulte souvent d'incohérences entre les annotations des ensembles de test et d'apprentissage. Nous montrons comment le principe de variation des annotations, introduit par Dickinson & Meurers (2003) pour identifier automatiquement les erreurs d'annotation, peut être utilisé pour identifier ces incohérences et évaluer leur impact sur les performances des analyseurs morpho-syntaxiques.

ABSTRACT

Evaluating Annotation Divergences in the UD Project

This work points out that the drop in performance often observed when applying a Part-of-Speech tagger to out-domain data may result from divergences between the annotation of train and test sets. We show how the annotation variation principle, introduced by (Dickinson & Meurers, 2003) to automatically identify errors in gold standard can be used to identify inconsistencies between annotations and to evaluate their impact on prediction performance.

MOTS-CLÉS : Erreur d'annotation, analyse morpho-syntaxique, adaptation au domaine.

KEYWORDS: Annotation error, PoS-tagging, domain adaptation.

1 Introduction

Les performances des analyseurs morpho-syntaxiques statistiques chutent de manière significative dès qu'ils sont utilisés pour étiqueter des phrases provenant d'un domaine différent de celui sur lequel ils ont été entrainés. Cette chute est généralement expliquée (Seddah *et al.*, 2012; Plank *et al.*, 2014; Bartenlian *et al.*, 2017) en supposant que la différence entre domaines se traduit par un changement de la distribution des vecteurs de caractéristiques $p(\mathbf{x})$ [1] liée à une variation dans l'usage des mots, à l'augmentation du nombre de mots hors-vocabulaire, à un usage différent de la ponctuation ou de la typographie, etc.

Cet article défend un point de vue différent et cherche à mettre en évidence qu'un autre facteur important pour expliquer la différence entre la entre domaines est le changement de la distribution jointe $p(\mathbf{x}, \mathbf{y})$, qui découle d'incohérences dans la manière dont les corpus sont étiquetés. En effet,

1. Cette situation est décrite dans la littérature statistique sous le nom de *covariate shift* (Shimodaira, 2000).

dans la quasi totalité des expériences d'adaptation au domaine, les ensembles de test et d'apprentissage ont été annotés de manière indépendante par des experts différents et même lorsqu'un même jeu d'étiquettes a été utilisé, les guides d'annotation ne sont pas toujours interprétés de la même manière. Le tableau 1 donne plusieurs exemples [2] de ces *divergences entre annotations*.

①	◇ Pour les particuliers , ce fut - et c' est toujours - une véritable aubaine , d' autant qu' en octobre 1989 , par peur d' une fuite de les placements dans une Europe bientôt sans frontières , le gouvernement socialiste (un comble !) exonéra pratiquement d'$_{\text{ADP}}$ impôts les revenus de les sicav monétaires , admises à le revenu de ces sicav à le voisinage de 10 % : ◇ Pour les particuliers , ce fut - et c' est toujours - une véritable aubaine , d' autant qu' en octobre 1989 , par peur d' une fuite de les placements dans une Europe bientôt sans frontières , le gouvernement socialiste (un comble !) exonéra pratiquement d'$_{\text{DET}}$ impôts les revenus de les sicav monétaires , admises à le revenu de ces sicav à le bénéfice de la capitalisation (voir ci - dessous) .
②	◇ Le Marché commun de le Golfe , devant déboucher vers la mise en place d' une monnaie commune , connaît aujourd'hui des retards en raison , entre autres$_{\text{NOUN}}$ de les divergences nées entre Ryadh et Abou Dhabi , principalement à le sujet de le siège de la future Banque$_{\text{PROPN}}$ centrale . ◇ Le Marché commun de le Golfe , devant déboucher vers la mise en place d' une monnaie commune , connaît aujourd'hui des retards en raison , entre autres$_{\text{PRON}}$ de les divergences nées entre Ryadh et Abou Dhabi , principalement à le sujet de le siège de la future Banque$_{\text{NOUN}}$ centrale .
③	◇ A quelques jours de le sommet de les sept grands pays industrialisés$_{\text{ADJ}}$, de le 6 à le 8 juillet à Munich , nous poursuivons la radioscopie de la situation économique de les pays riches (le monde de les 30 juin , 1 et 2 juillet) . ◇ A quelques jours de le sommet de les sept grands pays industrialisés$_{\text{VERB}}$, de le 6 à le 8 juillet à Munich , nous poursuivons notre enquête sur la situation de les pays riches (le monde de les 30 juin , 1 , 2 et 3 juillet) .

TABLE 1: Exemples de divergences entre les annotations des corpus français de l'UD : toutes ces phrases comportent une suite de mots en commun (les mots différents apparaissent dans une police plus claire) dont l'annotation est différente. Seules les étiquettes conflictuelles ont été représentées.

Dans cet article, nous décrivons plusieurs expériences qui mettent en évidence les divergences entre annotations au sein du projet *Universal Dependencies* (désormais UD)(Nivre *et al.*, 2017). Suivant l'approche développée par Boudin & Hernandez (2012), nous montrons comment le principe de variation d'annotations *(annotation variation principle)* introduit par Dickinson & Meurers (2003) peut être utilisé pour identifier les divergences entre annotations et mesurons l'impact de celles-ci sur l'évaluation des performances d'un analyseur morpho-syntaxique. Nos résultats suggèrent que la performance des étiqueteurs morpho-syntaxique sur des corpus hors domaine est souvent sous-estimée.

Le reste de cet article est organisé de la manière suivante. Nous commencerons par décrire les corpus et outils utilisés dans nos expériences (§2). Nous analyserons ensuite les résultats de deux expériences révélant des divergences d'annotations dans les corpus de l'UD (§3), avant de quantifier leur impact sur la qualité des prédictions (§4).

2 Cadre expérimental

Données Toutes les expériences présentées dans ce travail ont été réalisées avec les données du projet *Universal Dependencies* [3] (Nivre *et al.*, 2017). Ce projet a pour objectif de développer des

2. Les exemples sont présentés après application de pré-traitements, ce qui explique la présence de formes « décontractées » pour au (*à le*), et les espaces autour des symboles de ponctuation.

3. Nous avons utilisé la version 2.1 des données.

corpus étiquetés en PoS et en dépendances pour un large éventail de langues. La dernière version de l'UD rassemble 102 corpus couvrant 60 langues. Pour 22 langues, plusieurs corpus sont disponibles (jusqu'à 5 pour le français [4] et le tchèque) et il est possible de réaliser 63 expériences d'adaptation, chacune correspondant à l'apprentissage d'un étiqueteur sur un domaine pour ensuite l'exploiter sur un autre [5].

De nombreux corpus de l'UD résultent d'une conversion manuelle ou semi-automatique d'un corpus existant vers le schéma d'annotation de l'UD (Bosco *et al.*, 2013; Lipenkova & Souček, 2014). Dans la mesure où ces annotations ou conversions ont été réalisées indépendemment par différentes équipes, le risque d'incohérences et d'erreurs lors de l'interprétation des guides d'annotation est démultiplié ; plusieurs travaux (Vilares & Gómez-Rodríguez, 2017) ont ainsi récemment montré que de nombreux corpus d'une même langue ne sont pas étiqueté de manière cohérente. Un des principaux objectifs du présent article est de confirmer et de quantifier ces observations.

Analyseur morpho-syntaxique Nos expériences utilisent un analyseur morpho-syntaxique à base d'historique (Black *et al.*, 1992). Dans ces modèles, la prédiction d'une séquence d'étiquettes morpho-syntaxiques se réduit à une succession de problèmes de classification multi-classe : les étiquettes des mots de la phrase sont prédits l'une après l'autre par un perceptron moyenné. Nous utilisons un jeu de caractéristiques standard (Zhang & Nivre, 2011) : le mot courant, les mots suivants et précédents dans une fenêtre de taille deux, les deux dernières étiquettes prédites, etc. Ce modèle permet d'atteindre des performances proches de l'état de l'art tout en étant extrêmement rapide à entrainer, ce qui permet de multiplier les expériences.

Une description détaillée de ce modèle se trouve dans (Wisniewski *et al.*, 2014b,a).

Variation d'annotations Le principe de « variation d'annotations » (Boyd *et al.*, 2008) repose sur l'intuition que si deux séquences de mots identiques sont étiquetées de manière différente, il est fort probable qu'une de ces annotations contient une erreur. Dans ce travail, nous utiliserons ce principe pour détecter les divergences entre annotations de deux corpus.

Nous appelons *match* une séquence de mots qui apparait dans au moins deux phrases provenant de deux corpus différents et dont les annotations sont différentes. L'identification des matchs nécessite, dans un premier temps, de repérer toutes les séquences de mots identique dans deux corpus. Il s'agit d'une instance du problème de la plus longue sous-chaine répétée *(maximal repeat problem)* qui permet d'extraire efficacement (c.-à-d. avec une complexité en temps et en espace linéaire par rapport à longueur des deux corpus) tous les matchs à l'aide d'un arbre de suffixes généralisé (Gusfield, 1997).

Les matchs peuvent correspondre à des mots ou des groupes de mots qui sont effectivement ambigus, et pour lesquels la présence d'une divergence d'annotation est justifiée. Nous considérons deux heuristiques pour filtrer ces faux positifs. Tout d'abord, nous supposons que plus un match est long, plus il est vraisemblable qu'il résulte d'une incohérence ou d'une erreur d'annotation. Ainsi, le tableau 1 contient des matchs contenant plus de 10 mots qui sont tous imputables à des erreurs

4. Le projet contient notamment les conversions du *French Treebank* (Abeillé *et al.*, 2003) et du corpus Sequoia (Candito & Seddah, 2012) dans le formalisme UD ainsi que des corpus collectés spécifiquement comme ParTuT développé à l'université de Turin.

5. Dans 23 conditions, au moins un des corpus est trop petit pour entrainer un analyseur morpho-syntaxique et le corpus ne peut être utilisé que pour tester le modèle appris.

features	min.	max.	médiane	% best
words	66,9%	98,8%	89,7%	38,3%
labels	59,5%	98,9%	90,5%	48,3%
combi	67,0%	98,8%	90,0%	13,4%

TABLE 2: Précision moyenne obtenue sur les 63 conditions considérées par un classifieur cherchant à identifier à quel corpus appartient une phrase donnée.

d'annotation. Deuxièmement, avec l'heuristique *disjointe*, nous supposons qu'un match correspond à une ambiguïté naturelle lorsque les ensembles de ses étiquettes dans le premier corpus et dans le second corpus sont complètement disjoints et ne prenons pas ces cas en considération.

3 Divergence d'annotations dans les corpus de l'UD

L'objectif de cette section est de quantifier les divergences d'annotation dans l'UD. Au §3.1, nous utiliserons la divergence $\mathscr{H}$ pour caractériser les différences entre les corpus d'une même langue. Nous ferons ensuite le lien entre les erreurs de prédiction et les variations d'annotation (§3.2).

3.1 Caractérisation des différences entre corpus

Pour caractériser la différence entre deux corpus, nous utilisons la divergence $\mathscr{H}$ introduite par Ben-David *et al.* (2010) pour mesurer la similarité entre deux domaines (ou deux corpus). Cette mesure peut être estimée par le taux d'erreur d'un classifieur binaire entrainé pour décider si une phrase annotée provient du premier ou du second domaine. Intuitivement, plus ce taux d'erreur est élevé, plus il est difficile de discriminer les corpus d'apprentissage des corpus de test, et donc plus on peut penser qu'ils sont similaires.

Dans nos expériences, nous utilisons un modèle bayésien naïf [6] et trois ensembles de caractéristiques pour décrire une phrase et son annotation : *words*, pour lequel chaque exemple est représenté par un sac de mots (unigrammes et bigrammes) ; *labels*, dans lequel les exemples sont représentés de la même manière, mais en considérant cette fois, les PoS à la place des mots ; et *combi* qui utilise la même représentation après que les mots ont été concaténés avec leur PoS. Le premier jeu de caractéristiques permet d'identifier une différence dans la distribution des observations, les deux derniers des divergences d'annotation.

Le tableau 2 rapporte les résultats de cette expérience. Il indique, pour chaque jeu de caractéristiques considéré, pour quel pourcentage des 63 expériences d'adaptation possibles avec les corpus UD, ce jeu obtient le plus petit taux d'erreur. La figure 1 détaille ces résultats pour le français [7]. Les résultats sur les autres langues montrent des tendances similaires.

Il apparait que, dans la plupart des cas, il est possible d'identifier correctement le corpus d'où provient une phrase et son annotation. Bien que les chiffres bruts soient difficiles à interpréter (les

6. Nous avons utilisé l'implémentation fournie par (Pedregosa *et al.*, 2011) sans optimiser les hyper-paramètres.
7. Tous les scores sont moyennés sur 10 apprentissages et tests

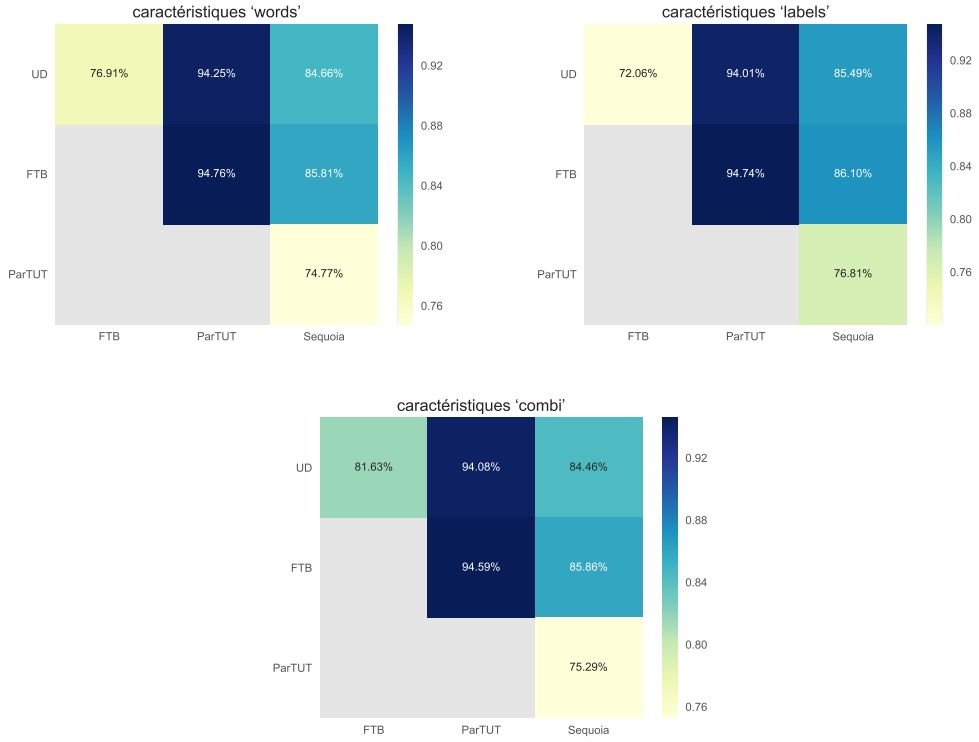

FIGURE 1: Précision obtenue par un classifieur cherchant à identifier à quel corpus appartient une phrase donnée sur les corpus français.

scores moyennés proviennent de nombreuses conditions expérimentales différentes, dont certaines correspondent à des problèmes de classification très déséquilibrés [8], ces résultats montrent clairement que, pour presque toutes les conditions, les caractéristiques les plus discriminantes incluent une description des étiquettes.

3.2 Impact des variations entre annotations sur la prédiction

Pour faire le lien entre erreurs de prédiction et divergences d'annotations, nous estimons le nombre de matchs dans lesquels au moins une étiquette n'est pas correctement prédite. En utilisant l'heuristique « disjointe » pour filtrer les matchs, il apparait que 70,2% (resp. 73,0%) des correspondances pour l'anglais (resp. français) contiennent une erreur de prédiction. Ces chiffres tombent à 51,7% (resp. 49,9%) lorsque les matchs ne sont pas filtrés et contiennent donc plus de mots ambigus. La figure 2 montre que filtrer les matchs selon leur longueur conduit à un effet similaire.

Toutes ces observations suggèrent que les variations entre annotations donnent souvent lieu à des erreurs de prédiction, surtout lorsqu'il est probable que la variation résulte d'une erreur d'annotation.

4 Évaluation hors domaine d'un analyseur morpho-syntaxique

Pour évaluer l'impact des erreurs d'annotation sur la qualité des prédictions d'un analyseur morpho-syntaxique, nous proposons de comparer $\varepsilon_{\text{full}}$, le taux d'erreur d'un analyseur estimé sur l'ensemble

8. Le rapport entre le nombre d'exemples de chaque corpus peut atteindre 88.

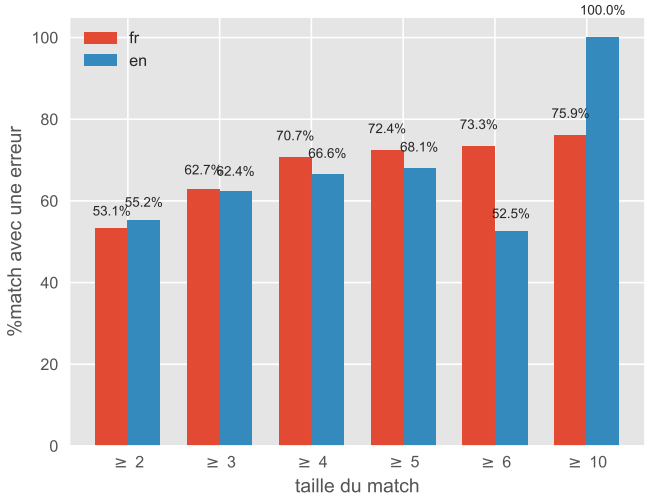

FIGURE 2: Matchs contenant au moins une étiquette mal prédite en fonction de la taille du match. Selon les jeux de données, le taux d'erreur sur l'ensemble du corpus est compris entre 5% et 15%.

du corpus de test et $\varepsilon_{\text{ignoring}}$ le taux d'erreur estimé en ignorant les matchs. Plus précisément, $\varepsilon_{\text{ignoring}}$ est défini de la manière suivante :

$$\varepsilon_{\text{ignoring}} = \frac{\#\{\text{errors}\} - \#\{\text{errors in matches}\}}{\#\{\text{words}\}} \tag{1}$$

où $\#\{\text{errors in matches}\}$ est le nombre d'erreurs dans des matchs, après filtrage. Intuitivement, ce taux d'erreur correspond à un taux d'erreur « oracle » qui serait obtenu si l'analyseur morpho-syntaxique prédisait correctement toutes les étiquettes des matchs.

La figure 3 représente ces taux d'erreur pour les corpus français lorsque sont ignorés les matchs ne respectant pas l'heuristique disjointe, les matchs de trois mots et plus, enfin tous les matchs. Supposer que les étiquettes des matchs sont correctement prédites permet de réduire de manière importante le taux d'erreur, au point que $\varepsilon_{\text{ignoring}}$ est souvent du même ordre de grandeur que le taux d'erreur obtenu sur un corpus du même domaine. En fait, dans plus de 43% (resp. 25%) des conditions, le taux d'erreur obtenu en ignorant les erreurs dans les matchs filtrés avec l'heuristique disjointe (resp. en fonction de la longueur du match) est inférieur au taux d'erreur obtenu sur les données du domaine.

Ces taux d'erreur sont naturellement sous-estimés puisque nous avons supprimé, dans ces expériences, des mots ou des structures qui étaient ambigus et dont l'étiquette est donc plus difficile à prédire. Ils peuvent toutefois être considérés comme une valeur oracle de la qualité de la prédiction.

Pour évaluer la qualité des différentes heuristiques de filtrage considérée, nous avons manuellement vérifié tous les matchs entre le corpus d'apprentissage `UD_French` et le corpus de test `FTB_French` et corrigé les différentes incohérences et erreur d'annotation (près de 2 000 étiquettes ont été modifiées). Lorsqu'il est entrainé sur le corpus d'origine, un analyseur morpho-syntaxique atteint un taux d'erreur de 6,8% sur le corpus FTB et 4,5% sur le corpus de test du même domaine. Après correction, le taux d'erreur hors-domaine tombe à 5,1%. Cette valeur est proche du taux d'erreur obtenu en ignorant les matchs de trois mots et plus, une observation qui valide l'heuristique considérée.

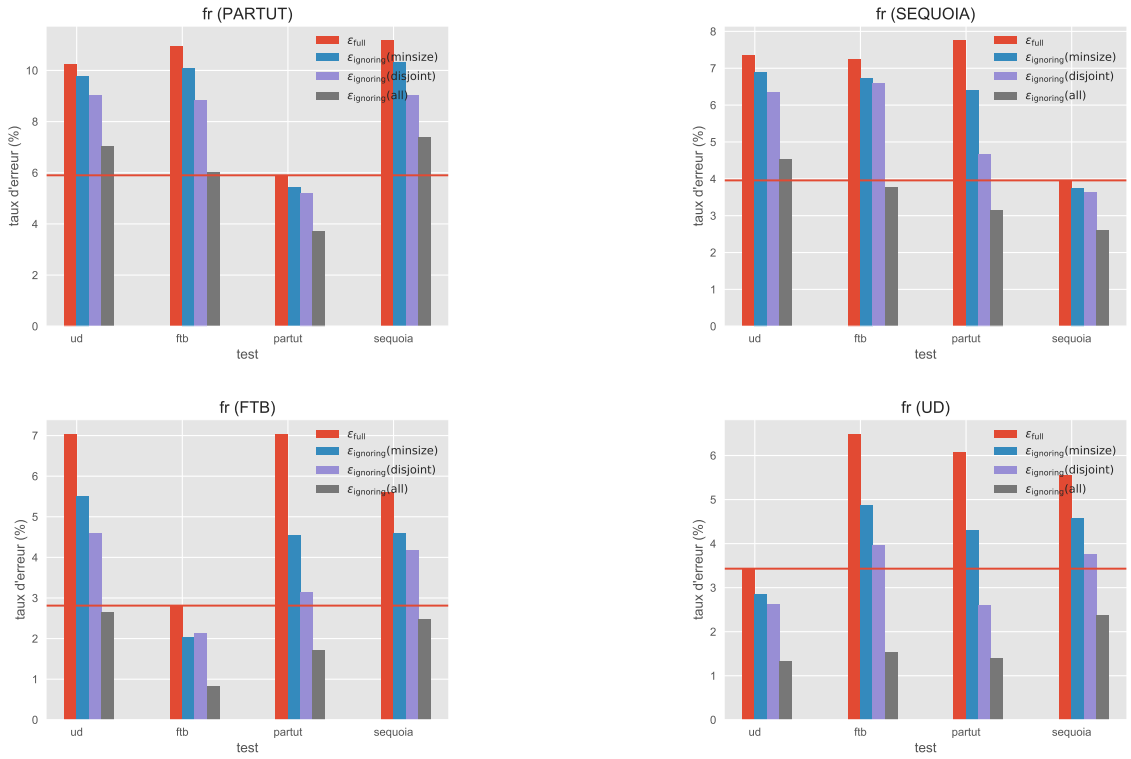

FIGURE 3: Taux d'erreur d'un analyseur morpho-syntaxique sur les différents corpus français de l'UD obtenus lorsque toutes les étiquettes des matchs sont correctement prédites. La ligne rouge représente le taux d'erreur obtenu sur un corpus de test du même domaine.

5 Conclusion

Dans ce travail, nous avons montré que, lors de l'évaluation d'un analyseur morpho-syntaxique sur des données hors-domaine, de nombreuses erreurs étaient dues à des divergence dans l'annotation des données. Une méthode permettant de quantifier cette divergence a également été décrite. Bien que nous n'ayons considéré que les corpus du projet UD et la tâche d'étiquetage morpho-syntaxique, la méthode décrite est très générique et peut être facilement applicable à d'autres corpus ou annotations (par exemple des annotations en dépendances), une tâche que nous aborderons dans nos futurs travaux.

Remerciements

Ces travaux ont été en partie financés par l'Agence Nationale de la Recherche (projet PARSITI, ANR-16-CE33-0021). Nous remercions les relecteurs pour leurs commentaires et suggestions.

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). *Building a Treebank for French*, In A. ABEILLÉ, Ed., *Treebanks : Building and Using Parsed Corpora*, p. 165–187. Springer Netherlands : Dordrecht.

BARTENLIAN E., LACOUR M., LABEAU M., ALLAUZEN A., WISNIEWSKI G. & YVON F. (2017). Adaptation au domaine pour l'analyse morpho-syntaxique. In *TALN 2017 - 24e conférence sur le Traitement Automatique des Langues Naturelles*, Orléan, France.

BEN-DAVID S., BLITZER J., CRAMMER K., KULESZA A., PEREIRA F. & VAUGHAN J. W. (2010). A theory of learning from different domains. *Machine Learning*, **79**(1-2), 151–175.

BLACK E., JELINEK F., LAFFERTY J., MAGERMAN D. M., MERCER R. & ROUKOS S. (1992). Towards history-based grammars : Using richer models for probabilistic parsing. In *Proceedings of the Workshop on Speech and Natural Language*, HLT'91, p. 134–139, Stroudsburg, PA, USA : Association for Computational Linguistics.

BOSCO C., MONTEMAGNI S. & SIMI M. (2013). Converting italian treebanks : Towards an Italian stanford dependency treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, p. 61–69, Sofia, Bulgaria : Association for Computational Linguistics.

BOUDIN F. & HERNANDEZ N. (2012). Détection et correction automatique d'erreurs d'annotation morpho-syntaxique du french treebank. In *TALN 2012 - 19e conférence sur le Traitement Automatique des Langues Naturelles*, p. 281–291, Grenoble, France.

BOYD A., DICKINSON M. & MEURERS W. D. (2008). On detecting errors in dependency treebanks. *Research on Language and Computation*, **6**(2), 113–137.

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

DICKINSON M. & MEURERS W. D. (2003). Detecting errors in part-of-speech annotation. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, p. 107–114, Stroudsburg, PA, USA : Association for Computational Linguistics.

GUSFIELD D. (1997). *Algorithms on Strings, Trees, and Sequences : Computer Science and Computational Biology*. New York, NY, USA : Cambridge University Press.

LIPENKOVA J. & SOUČEK M. (2014). Converting Russian dependency treebank to Stanford typed dependencies representation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2 : Short Papers*, p. 143–147, Gothenburg, Sweden : Association for Computational Linguistics.

NIVRE J., AGIĆ Ž., AHRENBERG L. & OTHER (2017). Universal dependencies 2.1. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

PLANK B., JOHANNSEN A. & SØGAARD A. (2014). Importance weighting and unsupervised domain adaptation of POS taggers : a negative result. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 968–973, Doha, Qatar : Association for Computational Linguistics.

SEDDAH D., SAGOT B., CANDITO M., MOUILLERON V. & COMBET V. (2012). The French Social Media Bank : a treebank of noisy user generated content. In *Proceedings of COLING 2012*, p. 2441–2458, Mumbai, India : The COLING 2012 Organizing Committee.

SHIMODAIRA H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, **90**(2), 227 – 244.

VILARES D. & GÓMEZ-RODRÍGUEZ C. (2017). A non-projective greedy dependency parser with bidirectional LSTMs. In *Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, p. 152–162, Vancouver, Canada : Association for Computational Linguistics.

WISNIEWSKI G., PÉCHEUX N., GAHBICHE-BRAHAM S. & YVON F. (2014a). Cross-lingual part-of-speech tagging through ambiguous learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1779–1785, Doha, Qatar : Association for Computational Linguistics.

WISNIEWSKI G., PÉCHEUX N., KNYAZEVA E., ALLAUZEN A. & YVON F. (2014b). Apprentissage partiellement supervisé d'un étiqueteur morpho-syntaxique par transfert cross-lingue. In *Proceedings of TALN 2014 (Volume 1 : Long Papers)*, p. 173–183, Marseille, France : Association pour le Traitement Automatique des Langues.

ZHANG Y. & NIVRE J. (2011). Transition-based Dependency Parsing with Rich Non-local Features. In *Proceedings of ACL 2011, the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, p. 188–193, Portland, Oregon, USA : Association for Computational Linguistics.