Combien d'exemples de tests sont-ils nécessaires à une évaluation fiable ? Quelques observations sur l'évaluation de l'analyse morpho-syntaxique du français.

Guillaume Wisniewski
LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91 405 Orsay, France
guillaume.wisniewski@limsi.fr

RÉSUMÉ

L'objectif de ce travail est de présenter plusieurs observations, sur l'évaluation des analyseurs morpho-syntaxique en français, visant à remettre en cause le cadre habituel de l'apprentissage statistique dans lequel les ensembles de test et d'apprentissage sont fixés arbitrairement et indépendemment du modèle considéré. Nous montrons qu'il est possible de considérer des ensembles de test plus petits que ceux généralement utilisés sans conséquences sur la qualité de l'évaluation. Les exemples ainsi « économisés » peuvent être utilisés en apprentissage pour améliorer les performances des systèmes notamment dans des tâches d'adaptation au domaine.

ABSTRACT

Some observations on the evaluation of PoS taggers

This work aims at reporting several observations on the evaluation of PoS taggers that are challenging the usual framework of statistical learning in which the test sets and are fixed arbitrarily and independently of the model considered. We show that, in many cases, it is possible to consider smaller test sets than those usually used with no impact on the quality of the evaluation.

MOTS-CLÉS : Apprentissage statistique, évaluation.

KEYWORDS: Machine Learning, Evaluation.

1 Introduction

L'apprentissage statistique est devenu la solution de choix pour la majorité des problèmes de traitement automatique des langues (TAL) : depuis que Charniak a montré que, pour l'analyse syntaxique, une méthode statistique surpassait les approches à base de règles (Charniak, 1996), on ne compte plus les tâches (identification de la polarité, reconnaissance d'entités nommées, traduction automatique, ...) pour lesquelles les meilleures performances (aussi bien évaluées automatiquement en comparant les sorties des systèmes à des références que par une évaluation qualitative réalisée à la main) sont obtenues en estimant les paramètres d'un modèle sur un corpus d'apprentissage.

La quasi totalité des travaux publiés aujourd'hui considère un même cadre expérimental pour évaluer aussi bien les modèles que les idées proposées : il existe, pour la plupart des tâches, des corpus de « références » (par exemple, pour l'analyse syntaxique, le Penn Tree Bank ou, pour la traduction automatique, les corpus des campagnes d'évaluation WMT) qui définissent un jeu de données d'apprentissage sur lesquels les modèles sont appris et un jeu de test réservé à l'évaluation des

performances de ceux-ci [1].

La répartition des données entre le corpus de test et le corpus d'apprentissage est généralement complètement arbitraire : des décisions aussi importantes que le nombre d'exemples que doivent contenir chacun de ces corpus reposent souvent sur des règles ou des savoir-faire empiriques voire complètement arbitraires [2] et ne sont que très rarement explicitées. Avoir des corpus de test et d'apprentissage fixes et clairement identifiés est supposé garantir facilement que les résultats publiés dans différents articles sont directement comparables et le jeu auquel nous jouons dans la plupart de nos publications consiste à améliorer les résultats de méthodes de références (les fameuses *baselines*) sur un ensemble de test donné et immuable.

Ce cadre expérimental ne correspond cependant pas à la plupart des applications « réelles » des méthodes de TAL : dans de nombreux cas, pour des raisons de coût ou de compétences, seules quelques dizaines voire quelques centaines d'exemples peuvent être étiquetées. Ceux-ci sont généralement réservés à l'évaluation des performances des modèles : constituer un ensemble d'apprentissage de taille suffisante (généralement plusieurs milliers voire dizaines de milliers d'exemples) est tout simplement irréaliste. C'est par exemple le cas lorsque l'on cherche à développer des modèles de TAL pour des langues peu dotées, notamment pour développer des outils d'aide à leur documentation (Michaud *et al.*, 2018), ou à des domaines pour lesquels il n'existe pas de données annotées (Sokolov *et al.*, 2017).

L'objectif de ce travail est de présenter plusieurs observations, sur une application particulière du TAL, l'analyse morpho-syntaxique du français, visant à remettre en cause le cadre habituel de l'apprentissage statistique dans lequel les ensembles de test et d'apprentissage sont fixés arbitrairement et indépendemment du modèle considéré. Les résultats présentés peuvent toutefois se généraliser facilement à tout système de TAL pouvant être évalué par un coût 0/1, Nous montrons que, bien souvent, il est possible de considérer des ensembles de test plus petits que ceux généralement définis sans conséquence sur la qualité de l'évaluation et que les exemples ainsi « économisés » peuvent être utilisés en apprentissage pour améliorer les performances des systèmes notamment dans des tâches d'adaptation au domaine.

2 Contexte

2.1 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 corpus étiquetés avec des informations morpho-syntaxiques pour un large éventail de langues. La dernière version de l'UD rassemble 133 corpus couvrant 75 langues. Le projet contient 7 corpus pour le français [4]. Ces corpus présentent une très grande variabilité dans les tailles des différents jeu de données utilisés : les corpus de test comportent entre 110 phrases (2 824 mots) et 2 541 phrases

1. Par soucis de clarté, nous ne parlerons pas, dans ce travail, des corpus de développement et supposerons qu'ils sont intégrés aux corpus d'apprentissage.

2. comme la règle bien connue : « 80% des données pour l'apprentissage et 20% pour le test ».

3. Nous avons utilisé la version 2.3 des données.

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.

	FTB	GSD	PUD	ParTUT	SRCMF	Sequoia	Spoken
Sequoia	92,4 $\pm$ 0,1	92,3 $\pm$ 0,5	86,3 $\pm$ 0,4	91,0 $\pm$ 1,1	52,3 $\pm$ 0,7	96,0 $\pm$ 0,3	80,5 $\pm$ 0,7
ParTUT	88,9 $\pm$ 0,2	89,2 $\pm$ 0,6	84,3 $\pm$ 0,4	94,2 $\pm$ 0,9	42,9 $\pm$ 0,7	88,6 $\pm$ 0,6	76,3 $\pm$ 0,8
GSD	93,2 $\pm$ 0,1	96,2 $\pm$ 0,3	89,7 $\pm$ 0,3	93,0 $\pm$ 1,0	54,4 $\pm$ 0,7	94,6 $\pm$ 0,4	83,8 $\pm$ 0,7
FTB	97,1 $\pm$ 0,1	93,1 $\pm$ 0,5	87,1 $\pm$ 0,4	93,5 $\pm$ 0,9	54,9 $\pm$ 0,7	94,4 $\pm$ 0,4	81,2 $\pm$ 0,7
Spoken	67,9 $\pm$ 0,3	69,4 $\pm$ 0,9	70,0 $\pm$ 0,5	74,6 $\pm$ 1,6	48,4 $\pm$ 0,7	70,4 $\pm$ 0,9	92,3 $\pm$ 0,5
SRCMF	60,8 $\pm$ 0,3	61,5 $\pm$ 0,9	63,4 $\pm$ 0,6	63,5 $\pm$ 1,8	92,5 $\pm$ 0,3	62,1 $\pm$ 0,9	65,1 $\pm$ 0,9

TABLE 1 – Précision (en %) d'un analyseur syntaxique appris et évalué sur les différentes combinaison d'ensemble de test et ensemble d'apprentissage des corpus français du projet Universal Dependencies. Les intervalles de confiance sont calculés en appliquant la méthode de Clopper-Pearson (c.f. § 2.2).

(82 440 mots), les corpus d'apprentissage entre 803 phrases (25 729 mots) et 14 759 phrases (485 464 mots). La table 1 rapporte les performances obtenues par un étiqueteur morpho-syntaxique utilisant un modèle à base d'historique (décrit dans le paragraphe suivant). Elle montre que, conformément à notre intuition, les performances chutent dès que l'on change de domaine et, surtout, que la confiance que l'on a dans l'estimation de la précision varie fortement selon les corpus.

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 prédites l'une après l'autre par un perceptron moyenné. Nous utilisons un jeu de caractéristiques standard (Zhang & Nivre, 2011). Une description détaillée de cet analyseur est faite dans (Bartenlian *et al.*, 2017; Wisniewski *et al.*, 2014b,a). Ce modèle permet d'atteindre des performances proches de l'état de l'art tout en étant extrêmement rapide à entraîner, ce qui permet de multiplier les expériences : il obtient une précision moyenne de 91,10% sur l'ensemble des corpus du projet UD (Aufrant *et al.*, 2017), un résultat comparable au 92.22% obtenu par l'analyseur UDPIPE (Straka & Straková, 2017) utilisé comme système de référence dans le défi *Multilingual Parsing from Raw Text to Universal Dependencies* organisé dans le cadre de CoNLL'17.

Nous utilisons également comme point de comparaison l'analyseur morpho-syntaxique du projet `CoreNLP` développé à Stanford. Cet analyseur repose sur un modèle à maximum d'entropie et utilise un ensemble de caractéristiques riches et des dépendances cycliques (Toutanvoa & Manning, 2000; Toutanova *et al.*, 2003) dont les paramètres ont été estimés sur un des corpus du projet UD [5].

2.2 Évaluation d'un classifieur

Un analyseur morpho-syntaxique peut être vu (c'est, en tout cas, de cette manière qu'il est évalué) comme un classifieur prédisant pour chaque mot en contexte, son étiquette morpho-syntaxique et évalué exactement comme un classifieur multi-classe. Nous rappelons rapidement dans ce paragraphe le principe de l'évaluation d'un tel classifieur et expliquons comment la notion d'*intervalle de confiance* permet de mesurer la qualité de cette évaluation.

La qualité d'un classifieur f, c'est-à-dire sa capacité à prédire l'étiquette associée à une observation donnée (un mot dans le cas d'un analyseur morpho-syntaxique) est naturellement évaluée par l'*erreur*

5. La documentation du projet ne précise ni la version ni le corpus qui a été utilisé en apprentissage

en généralisation :

$$e_g = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(f(x), y)\right] \tag{1}$$

où $\mathcal{D}$ est la distribution selon laquelle les données sont générées et ℓ la fonction de coût du problème. Cette erreur ne peut, bien évidemment, pas être calculée directement puisque ce calcul nécessiterait de connaître l'ensemble des données possibles et leur étiquette. Il est toutefois possible de l'estimer sur un échantillon $(x_i, y_i)_{i=1}^n$ de n exemples étiquetés :

$$\hat{e}_n = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i) \tag{2}$$

Si ces données n'ont pas été utilisées pour choisir les paramètres du classifieur, alors $\hat{e}_n$ est un estimateur non biaisé de l'erreur en généralisation (Duda *et al.*, 2001).

Les données utilisées pour estimer l'erreur en généralisation sont habituellement choisies à priori et indépendamment du classifieur et ce choix est fixe : pour la plupart des tâches existantes, la séparation entre ensemble de test et ensemble d'apprentissage est fixée par les personnes ayant collecté les données ou ayant défini la tâche et ce choix n'est quasiment jamais remis en question.

La valeur de $\hat{e}$ va naturellement dépendre de l'échantillon (c.-à-d. du choix du corpus de test) sur lequel elle est estimée et $\hat{e}$ peut être modélisé par une variable aléatoire distribuée selon une loi binomiale. Il est possible de caractériser la qualité d'une estimation, c'est-à-dire la différence entre la valeur réelle d'un paramètre (dans notre cas, l'erreur en généralisation) et son estimation (ici, l'erreur calculée sur un ensemble de test) en construisant un intervalle de confiance de niveau donné (Wasserman, 2013) qui définit une *marge d'erreur* entre la valeur estimée sur un échantillon et un relevé exhaustif sur la population totale.

Le niveau d'un intervalle de confiance, généralement exprimé sous la forme d'un pourcentage, minore la probabilité de contenir la valeur à estimer. Par exemple, si C est un intervalle de confiance à 95% du taux d'erreur, alors on sait que si on construit n intervalles de confiance de la même manière (par exemple en ré-échantillonnant l'ensemble de test) alors, pour n suffisamment grand, au moins 95% d'entre eux contiendront la « vraie » valeur du paramètre à estimer, c'est-à-dire l'erreur en généralisation.

Il est possible de construire un intervalle de confiance pour une variable aléatoire binomiale à l'aide de la méthode de Clopper-Pearson (Clopper & Pearson, 1934) dont il existe des implémentations pour la plupart des langages.

3 Quelques observations expérimentales

3.1 Impact de la taille de l'ensemble de test sur l'évaluation

Pour illustrer l'impact de la taille de l'ensemble de test sur la qualité de l'évaluation nous proposons de réaliser l'expérience suivante : les paramètres d'un analyseur morpho-syntaxique sont estimés sur l'ensemble d'apprentissage du corpus GSD ; les performances de cet analyseur sont ensuite estimées sur le corpus d'apprentissage du corpus FTB en considérant des corpus de test contenant un nombre d'exemples croissant. Utiliser un ensemble d'apprentissage (mais d'un corpus différent !) permet de garantir que l'on dispose de suffisamment de données (l'ensemble d'apprentissage contient 5 fois

plus de phrases que l'ensemble de test) pour mesurer l'impact de la taille de l'ensemble de test. Des résultats similaires à ceux que nous allons exposer dans la suite de cette section ont été obtenus sur les autres combinaisons d'ensemble de test et d'ensemble d'apprentissage. Par soucis de clarté nous ne détaillerons que les résultats obtenus sur les deux plus grands corpus du corpus UD.

Afin de caractériser l'influence du taux d'erreur, nous considérons, dans cette expérience, trois analyseurs différents : `gsd-full`, l'analyseur implémentant un modèle à base d'historique (c.f. section 2.1) appris sur la totalité du corpus d'apprentissage, `gsd-small` le même modèle appris sur la moitié du corpus d'apprentissage et`stanford`, l'analyseur de `CoreNLP` utilisant le modèle pré-entrainé fourni par les développeurs de cet outil.

La figure 1 représente l'évolution de la précision en fonction de la taille du corpus de test et surtout l'intervalle de confiance à 95% correspondant. Comme expliqué à la section 2.2, la largeur de l'intervalle de confiance permet de mesurer la qualité de l'estimation réalisée. La figure 2 montre l'évolution de cette largeur en fonction de la taille de l'ensemble de test.

Il apparaît que la largeur de l'intervalle de confiance diminue très vite avec le nombre de données et, en pratique, on pourrait réduire la taille du corpus d'évaluation de moitié sans impacter sensiblement la qualité de l'estimation et donc de l'évaluation. En effet, la largeur de l'intervalle de confiance estimé à partir de 40 000 mots est sensiblement le même que celle estimée à partir de 80 000 mots (correspondant, *grosso modo* à la taille du corpus de test « officiel ») : les intervalles de confiance sont, respectivement, $[93{,}8\%, 94{,}3\%]$ et $[93{,}6\%, 94{,}0\%]$. En outre, considérer encore plus d'exemples (p. ex. 160 000) ne permet de réduire la largeur de l'intervalle de confiance que de $0{,}1$ point.

Cette observation est renforcée par le fait qu'à partir d'un corpus de test comportant 40 000 mots, les intervalles de confiance des différents analyseurs ne se chevauchent plus et que la différence entre leurs performances est donc statistiquement significative (Wasserman, 2013). Ainsi, indépendamment de la largeur réelle de l'intervalle de confiance, les évaluations réalisées sont suffisamment précises pour permettre de répondre de manière convaincante à l'une des principales motivations de l'évaluation d'un modèle de TAL : est-ce qu'un analyseur améliore les résultats de la *baseline* ?

3.2 Conséquences pour les problèmes d'adaptation au domaine

Nous avons montré, dans la section précédente, qu'il était possible de réduire la taille du corpus de test tout en continuant d'évaluer la qualité d'un modèle avec suffisamment de précision pour pouvoir comparer deux modèles de manière fiable. Nous souhaitons illustrer dans cette section, les conséquences de cette observation sur la tâche d'adaptation au domaine pour l'analyse morpho-syntaxique.

Le cadre expérimental généralement considéré dans les tâches d'adaptation au domaine est le suivant : on dispose d'un corpus de données étiquetées suffisamment grand pour permettre l'apprentissage d'un modèle et d'un corpus de test contenant des données issues d'un domaine différent et ne permettant que d'évaluer la dégradation des performances liées au changement de domaine. C'est, par exemple, ce cadre qui est mis en œuvre dans les expériences décrites dans la table 1.

Motivés par les résultats présentés dans la section précédente, nous souhaitons évaluer l'hypothèse suivante : il est possible de continuer à évaluer les performances d'un analyseur morpho-syntaxique hors-domaine en considérant un corpus de test plus petit que ceux habituellement considérés et d'utiliser les exemples ainsi « économisés » en apprentissage.

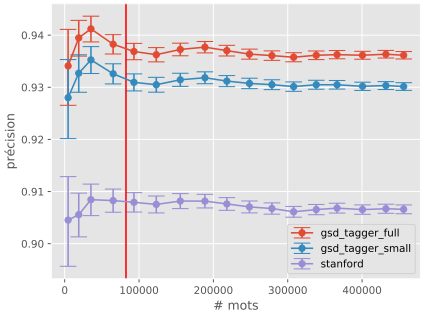

FIGURE 1 – Précision obtenue par 3 analyseurs morpho-syntaxiques estimée sur des corpus de tests de différentes tailles et les intervalles de confiance (à 95%) déterminés par la méthode de Clopper-Pearson correspondant. La ligne rouge correspond à la taille de l'ensemble de test « officiel ».

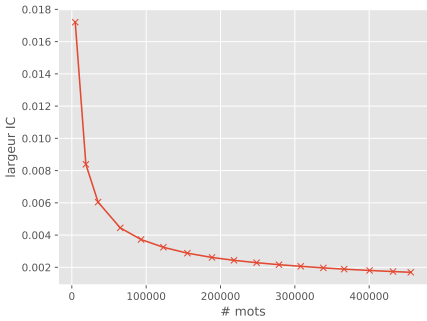

FIGURE 2 – Évolution de la marge d'erreur (largeur de l'intervalle de confiance à 95%) en fonction de la taille du corpus de test.

La figure 3 illustre ce principe. Elle représente l'évolution de l'erreur sur le corpus de test en fonction de la taille de celui-ci pour quatre modèles :

- **GSD** : un modèle à base d'historique appris uniquement sur les données d'apprentissage du corpus GSD (évaluation *out-domain*) ;
- **FTB** un modèle à base d'historique appris uniquement sur les données d'apprentissage sur la totalité du corpus FTB (évaluation *in-domain*) ;
- **FTB-small** un modèle à base d'historique appris uniquement sur les données d'apprentissage issues du corpus FTB et « économisé » en réduisant la taille du corpus d'évaluation ;
- **GSD+FTB-small** un modèle à base d'historique appris sur la concaténation des données GSD et des données économisées en test.

Les résultats présentés à la figure 3 montrent qu'il est possible d'améliorer significativement (les intervalles de confiance ne se chevauchent pas !) les performances d'un analyseurs morpho-syntaxique hors-domaine en utilisant une partie des données étiquetées disponibles pour l'apprentissage plutôt que pour l'évaluation et ce, sans impact sur la confiance que l'on a dans l'estimation de la qualité d'un système.

4 Conclusions

Nous avons décrit, dans ce travail, plusieurs observations sur l'évaluation d'un système d'analyse morpho-syntaxiques du français. Ces observations montrent qu'une connaissance, à priori, de l'ordre de grandeur du taux d'erreur permet de réduire de manière significative le nombre d'exemples nécessaires à l'évaluation de l'erreur de généralisation. Même si nous n'avons considéré dans nos expériences que la tâche d'analyse morpho-syntaxique, l'approche décrite peut être appliquée sans problème à toute tâche s'évaluant avec un coût 0/1. Nos travaux futurs porteront, entre autres, sur la possibilité de généraliser notre approche à d'autres fonctions de coût.

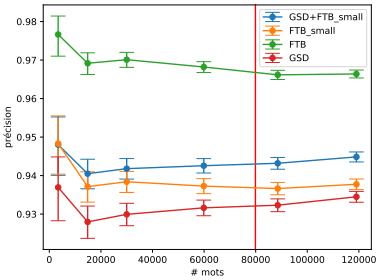

FIGURE 3 – Précision obtenue par un analyseur morpho-syntaxique dans un cadre « adaptation au domaine » en fonction de la taille de l'ensemble d'évaluation.

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.

AUFRANT L., WISNIEWSKI G. & YVON F. (2017). LIMSI@CoNLL'17 : UD shared task. In *Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, p. 163–173, Vancouver, Canada : Association for Computational Linguistics.

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.

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.

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.

CHARNIAK E. (1996). *Tree-bank Grammars*. Rapport interne, Providence, RI, USA.

CLOPPER C. J. & PEARSON E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, **26**(4), 404–413.

DUDA R. O., HART P. E. & STORK D. G. (2001). *Pattern Classification*. New York : Wiley, 2 edition.

MICHAUD A., ADAMS O., COHN T. A., NEUBIG G. & GUILLAUME S. (2018). Integrating automatic transcription into the language documentation workflow : Experiments with na data and the persephone toolkit. *Language Documentation & Conservation*, p. 393–429.

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.

SOKOLOV A., KREUTZER J., SUNDERLAND K., DANCHENKO P., SZYMANIAK W., FÜRSTENAU H. & RIEZLER S. (2017). A shared task on bandit learning for machine translation. In *Proceedings of the Second Conference on Machine Translation*, p. 514–524 : Association for Computational Linguistics.

STRAKA M. & STRAKOVÁ J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, p. 88–99, Vancouver, Canada : Association for Computational Linguistics.

TOUTANOVA K., KLEIN D., MANNING C. D. & SINGER Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

TOUTANVOA K. & MANNING C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

WASSERMAN L. (2013). *All of Statistics*. Springer.

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.