Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Régression logistique parcimonieuse pour l'extraction automatique de règles de grammaire

Santiago Herrera[1] Caio Corro[2] Sylvain Kahane[1,3]
(1) Université Paris Nanterre, CNRS, Modyco, 200 Avenue de la République, 92001, Nanterre, France
(2) Sorbonne Université, CNRS, ISIR, 4 Place Jussieu, 75005, Paris, France
(3) Institut Universitaire de France
s.herrera@parisnanterre.fr, caio.corro@isir.upmc.fr, sylvain@kahane.fr

RÉSUMÉ

Nous proposons une nouvelle approche pour extraire et explorer des motifs grammaticaux à partir de corpus arborés, dans le but de construire des règles de grammaire syntaxique. Plus précisément, nous nous intéressons à deux phénomènes linguistiques, l'accord et l'ordre des mots, en utilisant un espace de recherche étendu et en accordant une attention particulière au classement des règles. Pour cela, nous utilisons un classifieur linéaire entraîné avec une pénalisation L1 pour identifier les caractéristiques les plus saillantes. Nous associons ensuite des informations quantitatives à chaque règle. Notre méthode permet de découvrir des règles de différentes granularités, certaines connues et d'autres moins. Dans ce travail, nous nous intéressons aux règles issues d'un corpus du français.

ABSTRACT

Sparse Logistic Regression with High-order Features for Automatic Grammar Rule Extraction from Treebanks

We propose a novel approach to extract and explore significant fine-grained grammar patterns and potential syntactic grammar rules from treebanks, in order to create an easy-to-understand corpus-based grammar. More specifically, we extract descriptions and rules across different languages for two linguistic phenomena, agreement and word order, using a large search space and paying special attention to the ranking order of the extracted rules. For that, we use a sparse logistic regression classifier to extract the most salient features that predict the linguistic phenomena under study. We associate statistical information to each rule. Our method discovers both known and significant grammar rules that are less well known. In this paper, we focus on rules extracted from a French treebank.

MOTS-CLÉS : Extraction de grammaire, règles de grammaire, grammaire fondée sur des corpus, grammaire quantitative, régression logistique, pénalité L1.

KEYWORDS: Grammar extraction, grammar rules, corpus based grammar, quantitative grammar, sparse logistic, L1 regularization.

1 Introduction

Les grammaires descriptives sont des ressources précieuses pour la recherche linguistique, mais leur construction est longue et difficile. Les collections de corpus arborés comme Universal Dependencies

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 211–218.

(De Marneffe *et al.*, 2021) ont permis le développement de méthodes d'extraction automatique de grammaires à partir de corpus. Dans ce travail, nous proposons une nouvelle approche fondée sur la régression logistique parcimonieuse. Nos contributions peuvent se résumer de la façon suivante :

— nous proposons une nouvelle formalisation des règles grammaticales visant l'extraction automatique à partir de corpus arborés ;
— nous étudions l'utilisation de la régression logistique L1 pour extraire et classer les règles ;
— par rapport aux travaux précédents (notamment Chaudhary *et al.* 2020, 2022), nous augmentons l'expressivité de nos règles, ce qui permet d'obtenir des règles plus fines et d'avoir aussi un outil plus adapté à la fouille de règles

Nous nous évaluons sur le français, l'espagnol et wolof pour trois phénomènes : les accords en genre, en nombre et l'ordre des mots. Cependant, nous ne reportons que les résultats sur la position du sujet en français dans ce résumé. [1]

2 Règle de grammaire

Une grammaire est un ensemble de contraintes régulières qu'une langue impose à ses locuteurs, contraintes que nous appelons souvent des règles. Une règle de grammaire décrit le contexte linguistique particulier dans lequel un motif grammatical est privilégié ou non, par rapport à d'autres motifs possibles. Dans la pratique, ces règles sont de nature probabiliste (dans l'interprétation fréquentiste), peuvent être plus ou moins fines et peuvent se chevaucher.

Par exemple, une règle simple du français est : « *le sujet nominal d'un verbe se place avant son gouverneur* ». Cependant, cette règle n'est pas déterministe. Une description plus précise serait : « *pour une dépendance syntaxique de type* sujet*, le dépendant se place avant son gouverneur dans plus de* 97% *des cas* » [2]. En pratique, il est intéressant de comprendre dans quels cas il y a une divergence avec la règle dominante, par exemple : « *le sujet nominal dans une subordonnée relative se place après son gouverneur dans* $\approx 23\,\%$ *des cas* » (voir la Figure 1).

Nous proposons de définir une règle de grammaire comme une construction composée de trois motifs :

— S est un motif donné qui définit la portée (angl. *scope*) de la règle, c'est-à-dire quelles sont les constructions qui nous intéressent (p. ex. nous pouvons nous focaliser sur les relations de type *sujet* uniquement).
— Q identifie le phénomène ou la question linguistique qui nous intéresse (p. ex. l'inversion du sujet).
— P est un motif prédicteur ou déclencheur de Q dans les limites de S (p. ex. le fait que le sujet soit dans une subordonnée relative).

Nous formalisons donc une règle de grammaire par une formule de la forme suivante :

$$\mathrm{S} \implies (\mathrm{P} \overset{\alpha\,\%}{\Longrightarrow} \mathrm{Q})\,.$$

où α est la fréquence avec laquelle le motif P déclenche le motif Q au sein de S. Une reformulation naturelle de notre exemple serait « parmi tous les sujets, ceux qui se trouvent dans une proposition

1. Une version en anglais de ce travail a été accepté à la conférence LREC-Coling 2024 : https://arxiv.org/abs/2403.17534.
2. Distribution tirée de la version SUD du corpus arborés GSD du français (Guillaume *et al.*, 2019).

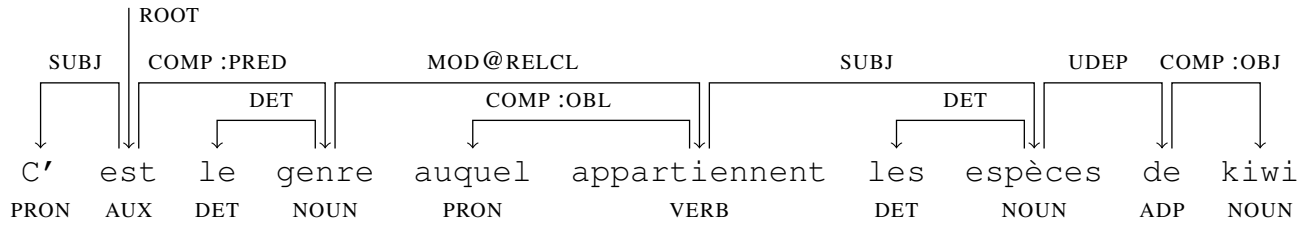

FIGURE 1 – Exemple tiré de la version SUD du corpus arborés GSD du français. Deux propositions sont reliées par une relation MOD@RELCL avec des sujets dans des positions différentes. Dans la première (resp. seconde) proposition, le sujet occupe une position préverbale (resp. post-verbale).

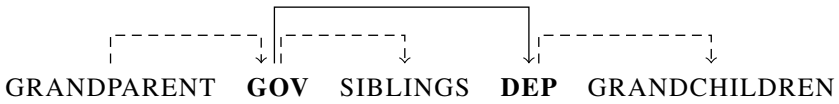

FIGURE 2 – L'espace de recherche est définit autour d'une relation entre un nœud gouverneur (GOV), et son dépendant (DEP), le gouverneur de GOV (GRANDPARENT) et les autres dépendants de GOV (SIBLINGS), ainsi que les dépendants de DEP (GRANDCHILDREN). On utilise presque toutes les informations linguistiques encodées dans ces nœuds, y compris les traits morphosyntaxiques, certains lemmes et les relations de dépendances, ainsi que les positions relatives de ces nœuds (qui peuvent être différentes de celles de la figure).

relative sont inversés dans plus de 23% des cas ».

Cette formalisation s'inspire des règles (de correspondance) de la théorie Sens-Texte (Mel'čuk, 1988) [3]. Cependant, nous nous concentrons uniquement sur les relations au niveau de la syntaxe de surface. De plus, cette formalisation se traduit facilement en une tâche de classification par apprentissage automatique.

En pratique, les motifs S et Q sont déterminés manuellement, car ils définissent les phénomènes linguistiques auxquels nous allons nous intéresser. Cependant, le nombre potentiel de motifs P est très important et il est donc difficile de les explorer manuellement. En particulier, nous utilisons presque toutes les informations linguistiques encodées dans les corpus arborés : pour toutes les dépendances filtrées par S, notre espace de recherche pour P considère toute information relative au grand-parent (gouverneur du gouverneur), aux codépendants du gouverneur et aux enfants du dépendant (voir Figure 2).

3 Méthode d'extraction

Plusieurs études ont démontré la nécessité d'avoir une approche quantitative pour étudier des phénomènes syntaxiques (p. ex. Bresnan *et al.*, 2004; Thuilier, 2012). Aujourd'hui, il est possible d'extraire automatiquement des règles syntaxiques en utilisant des corpus annotés en syntaxe comme *Universal Dependencies* (UD, De Marneffe *et al.*, 2021) ou *Surface Syntactic UD* (SUD, Gerdes *et al.*, 2018). Les travaux de Chaudhary *et al.* (2020, 2022) proposent d'utiliser des arbres de décision,

3. Dans la notation de Mel'čuk (1988), la règle s'écrirait « S $\Longrightarrow$ Q | P » et se lirait « S est susceptible de correspondre à Q dans le contexte P » (le sujet peut être placé après le verbe quand il est dans une relative).

comme technique d'apprentissage automatique, à cette fin. Nous nous différencions de ces travaux en proposant une nouvelle formalisation du problème (section précédente) et car nous utilisons la régression logistique parcimonieuse, qui est plus sensible à des changements même peu prononcés de la distribution. Cela nous permet de simultanément extraire et classer les règles, tout en explorant un large espace de recherche. Nous obtenons ainsi des règles de grammaire quantitative plus expressives et fines. De plus, notre approche ne nécessite pas de régler des hyperparamètres d'apprentissage.

3.1 Régression logistique et pénalisation L1

Dans cette section, nous décrivons brièvement notre méthode d'extraction des motifs P fondée sur la régression logistique parcimonieuse. Nous renvoyons les lecteurices vers Shalev-Shwartz & Ben-David (2014, Sec. 9 & 25) et Bach *et al.* (2012) pour une introduction détaillée à ce sujet.

Modèle de classification. Soit F l'ensemble des motifs possibles pour P, qui correspondent aux caractéristiques d'entrée de notre classifieur. Nous considérons comme caractéristiques tous les motifs composés d'un ou de deux attributs dans l'espace de recherche de la Figure 2, par exemple GRANDCHILDREN.UPOS=ADP et GOV.REL_SYNT=MOD,SIBLINGS.UPOS=ADP. Soit $\boldsymbol{x} \in \{0,1\}^F$ un vecteur booléen qui indique quelles sont les caractéristiques présentent dans l'entrée. Notre modèle de classification définie la probabilité d'observer un certain motif Q, par ex. un sujet inversé, de la façon suivante :

$$\text{P("inversion du sujet"}|\boldsymbol{x}) = \sigma\left(\boldsymbol{a}^\top \boldsymbol{x} + b\right) \tag{1}$$

où $\boldsymbol{a} \in \mathbb{R}^F$ et $b \in \mathbb{R}$ sont les paramètres du modèle, et $\sigma(w) = \frac{\exp(w)}{1+\exp(w)}$ est la fonction sigmoïde. Par construction du modèle, les caractéristiques qui ont un poids 0 dans le vecteur $\boldsymbol{a}$ ne contribuent pas à la distribution du modèle.

Nous apprenons les paramètres du modèle en minimisant l'objectif suivant :

$$\min_{\boldsymbol{a}\in\mathbb{R}^F, b\in\mathbb{R}} \quad \frac{1}{|D|} \sum_{(\boldsymbol{x},y)\in D} \ell\left(\boldsymbol{a}^\top \boldsymbol{x} + b, y\right) \quad + \quad \lambda r(\boldsymbol{a})\,, \tag{2}$$

où D est le corpus d'entraînement filtré par S, ℓ est la fonction de perte, r est une pénalité de régularisation de poids $\lambda \geq 0$. En particulier, nous utilisons :

- $\ell(w, y) = -yw + \log(1 + \exp w)$, l'opposé de la log-vraisemblance ;
- $r(\boldsymbol{a}) = \|\boldsymbol{a}\|_1 = \sum_{f\in F} |a_f|$, la norme L1 qui favorise les solutions contenant des 0 dans le vecteur $\boldsymbol{a}$ (Bach *et al.*, 2012).

Nous utilisons la bibliothèque SKGLM (Bertrand *et al.*, 2022) pour apprendre les paramètres.

Chemin de régularisation. Plus le paramètre λ est élevé, plus le nombre de 0 dans le vecteur $\boldsymbol{a}$ sera élevé. Nous entraînons donc une séquence de modèles en faisant varier le poids de la régularisation L1 afin d'identifier les motifs P du plus important au moins important. Nous initialisons λ à une valeur assez grande pour que toutes les valeurs dans $\boldsymbol{a}$ soit nulles. Ensuite, nous entraînons la séquence de modèles en diminuant petit à petit la valeur de λ. Plus un motif apparaît tôt avec un score non nul, plus il est considéré comme important. La séquence de solutions obtenues en diminuant la valeur de λ est appelée le chemin de régularisation (Markowitz, 1952; Osborne *et al.*, 2000; Efron *et al.*, 2004).

3.2 Filtrage et analyse des règles

Bien que le chemin de régularisation nous donne un classement des motifs P, les coefficients de $\boldsymbol{a}$ ne constituent pas en soi des facteurs explicatifs, contrairement aux idées reçues (Achen, 2005). C'est pour cela que, dans un premier temps, pour chaque motif extrait, nous vérifions s'il est déclencheur du motif Q ou ¬Q, autrement dit, si le motif fait pencher la distribution d'un côté ou de l'autre.

Dans un deuxième temps, nous calculons la précision et la couverture/rappel de la règle, et nous appliquons un test statistique pour mieux analyser les règles. La précision et la couverture de la règle sont particulièrement importantes, car nous faisons l'hypothèse qu'une règle de grammaire fiable couvre largement le phénomène linguistique d'intérêt (ex. l'inversion du sujet) et qu'elle comporte peu de faux positifs.

Règle	**Couverture**	**Précision**
$S \wedge P \overset{\alpha\,\%}{\Longrightarrow} Q$	$\frac{\#(S \wedge P \wedge Q)}{\#(S \wedge Q)}$	$\frac{\#(S \wedge P \wedge Q)}{\#(S \wedge P)}$
$S \wedge P \overset{\alpha\,\%}{\Longrightarrow} \neg Q$	$\frac{\#(S \wedge P \wedge \neg Q)}{\#(S \wedge \neg Q)}$	$\frac{\#(S \wedge P \wedge \neg Q)}{\#(S \wedge P)}$

L'utilisation de tests et de mesures statistiques permet de pondérer les règles selon différentes mesures et ainsi d'avoir une grammaire quantitative fondée sur les occurrences d'un corpus. Il convient de noter que ces mesures doivent être contextualisées. Une règle comme celle de l'inversion du sujet dans une relative n'atteindra jamais une couverture de $100\,\%$ (et sera probablement bien inférieure).

4 Résultats pour l'inversion du sujet

Toutes les expériences utilisent les corpus arborés SUD dans sa version 2.13. Notre méthode étant indépendante des langues étudiées, nous avons extrait des règles d'accord et d'ordre pour plusieurs langues ayant un treebank en dépendance. Dans ce résumé, nous nous limitons à exposer quelques règles positives pour l'ordre du sujet en français (Table 1) qui servent à évaluer nos résultats. Notre méthode peut contribuer à la recherche de règles pour des langues peu décrites, mais aussi pour des langues bien décrites où certaines généralisations ou certains cas particuliers ont pu nous échapper. Le code pour reproduire les expériences est disponible en ligne [4] ainsi qu'un outil permettant de visualiser les résultats sur différents corpus [5].

La règle la plus saillante pour l'inversion du sujet, ainsi que la deuxième et quatrième, indiquent que les sujets dans les incises (PARATAXIS:INSERT), dans une structure de discours rapporté (ex : "La route sera longue, prévient le représentant du pape."), sont toujours inversés. Bien entendu, ces règles montrent que l'extraction dépend de ce qui a été annoté : le treebank du français utilisé mentionne explicitement les propositions incises, en les distinguant des autres cas de parataxe.

La cinquième règle capture l'exemple privilégié dans ce résumé : être dans une relative (GOV.DEEP_REL=MOD@RELCL) déclenche à un certain degré l'inversion du sujet. On sait qu'il s'agit d'un sujet nominal car il est déterminé (GRANDCHILDREN.REL_SYNT=DET). Avoir capturé

4. https://github.com/FilippoC/grex-lrec-coling-2024
5. https://autogramm.github.io/grex-lrec-coling-2024

P de la règle	λ	couverture	précision
gov.rel_synt=parataxis:insert,grandparent.position=before_gov	0,009	24,9	100
gov.rel_synt=parataxis:insert	0,007	24,9	100
grandchildren.rels_synt=det,grandparent.position=before_gov	0,005	34,1	9,7
gov.VerbForm=Fin,gov.rel_synt=parataxis:insert	0,005	24,9	100
gov.rel_deep=mod@relcl,grandchildren.rels_synt=det	0,004	16,8	23,7
gov.VerbForm=Fin,siblings.upos=PRON	0,004	26,3	5,4
dep.rel_deep=subj@expl,gov.Person=3	0,003	13,2	10,4
dep.Person=3,dep.rel_deep=subj@expl	0,003	13,2	10,4
gov.VerbForm=Fin,grandparent.Number=Sing	0,003	38,7	7,4
grandchildren.rels_synt=det,siblings.upos=PRON	0,003	17,8	9,5

TABLE 1 – Les dix règles d'ordre des mots les plus saillantes qui favorisent l'inversion du sujet par rapport à son gouverneur, extraites de la version 2.13 du treebank SUD GSD du français et classées en fonction de l'ordre donné par le classifieur linéaire. Bien que les règles négatives (qui prédisent l'ordre canonique sujet-verbe) sont intéressantes, nous incluons seulement les règles positives. Voir la section 3.1 pour l'interprétation de λ. La couverture et la précision sont exprimées en pourcentage.

cette règle représente un bon test de l'expressivité de notre méthode étant donné sa faible précision (23,7%) et couverture (16,8%).

La troisième règle, que nous avions préalablement sautée, englobe celles déjà mentionnées. Elle comprend un motif P très général signalant une tendance syntaxique : le sujet déterminé se place plus souvent après son gouverneur par rapport à la distribution positionelle de base dans des propositions subordonnées. La subordination est ici encodée par l'existence d'un gouverneur du verbe (grandparent.position=before_gov).

La sixième et la dixième règle correspondent au cas des sujets où le verbe possède un dépendant pronominal. Il s'agit en fait d'une situation où prédomine les relatives et les interrogatives (ex : "Que dit l'Église sur ce point délicat ?"). Les règles présentent néanmoins une formulation inhabituelle et inattendue.

La septième règle, comme la huitième, capture le cas des sujets explétifs (SUBJ@EXPL), qui se trouvent être effectivement plus souvent post-verbaux dans les interrogatives (ex. "Ces chiffres sont-ils élevés ?" ; "Qu'est-ce qui va augmenter ?).

On aura noté qu'on obtient souvent des règles similaires déclenchées par des motifs qui se recoupent. Autrement dit, nous obtenons une hiérarchie de résultats plus ou moins fins. Il serait possible de repérer de tels motifs et de les filtrer, mais il n'est pas forcément évident de décider automatiquement qu'elle est la formulation la plus naturelle pour l'utilisateur et une étude plus approfondie est nécessaire. À notre connaissance, les règles détaillées n'ont pas été capturées par des travaux précédents (Chaudhary *et al.*, 2022).

Outre l'intérêt descriptif, les règles extraites ont aussi un intérêt comparatif. Il serait envisageable de réaliser des études contrastives entre plusieurs langues ou entre plusieurs corpus pour repérer les motifs saillants d'une langue ou d'un corpus par rapport à d'autres. La comparaison de motifs entre plusieurs langues ou familles de langues permettrait aussi de faire de la typologie quantitative avec un niveau de description plus fin que d'habitude. Par exemple, il rendrait possible non seulement de comparer l'ordre des constituants majeurs, mais aussi de comparer la distribution complète des sujets

dans une famille de langue. Ces expériences prometteuses requièrent des études spécifiques et sont hors portée de ce travail.

5 Limitations

Les règles extraites dépendent du schéma d'annotation, de la nature du corpus et de sa qualité. Ainsi, quelques motifs extraits expriment des propriétés du corpus ou des décisions théoriques, plutôt que des règles de grammaire à proprement parler. Nous sommes aussi limités par ce qui est effectivement annoté et nos règles sont donc circonscrites à la syntaxe de surface, c'est-à-dire aux règles de bonne formation de l'arbre syntaxique (règles d'accord) et aux règles de correspondance entre l'arbre et la chaîne linéaire (règle d'ordre des mots).

En outre, certains résultats observés montrent des préférences d'usage de la langue, et non des règles de grammaire. Par exemple, on note une tendance à l'accord du verbe avec son objet, du fait que les objets singuliers sont significativement appariés à des sujets singuliers (de même pour les pluriels). À l'inverse, la règle d'accord d'un participe passé avec son objet lorsque celui-ci le précède n'a pas été retrouvée, probablement parce que le pronom relatif *que* ne comporte pas de trait de nombre et de genre dans l'annotation (ce qui est par ailleurs un choix d'annotation tout à fait raisonnable si on s'en tient à l'annotation morphosyntaxique).

6 Conclusion

Nous proposons une nouvelle méthode d'extraction de règles grammaticales à partir de corpus arborés. Notre approche est fondée sur (1) une définition formelle de ce qu'est une règle grammaticale syntaxique et (2) l'utilisation d'un modèle linéaire de classification pour extraire et classer les règles via le chemin de régularisation. De plus, nous avons réussi à montrer avec notre analyse qu'il est important d'étendre l'espace de recherche pour augmenter l'expressivité des règles et pour capturer des phénomènes linguistiques qui sont hors de portée des travaux précédents (Chaudhary *et al.*, 2020, 2022; Blache *et al.*, 2016). Notre méthode est également mieux adaptée à la fouille de règles.

Nous espérons aussi, avec ce travail, contribuer à l'élaboration de grammaires descriptives et aussi de contribuer au rapprochement entre le TAL ou la Linguistique Computationnelle (CL) et la linguistique théorique, ainsi qu'entre le TAL/CL et la linguistique de terrain.

Références

ACHEN C. H. (2005). Let's put garbage-can regressions and garbage-can probits where they belong. *Conflict Management and Peace Science*, **22**(4), 327–339.

BACH F., JENATTON R., MAIRAL J., OBOZINSKI G. *et al.* (2012). Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, **4**(1), 1–106.

BERTRAND Q., KLOPFENSTEIN Q., BANNIER P.-A., GIDEL G. & MASSIAS M. (2022). Beyond l1 : Faster and better sparse models with skglm. In *NeurIPS*.

BLACHE P., RAUZY S. & MONTCHEUIL G. (2016). MarsaGram : an excursion in the forests of parsing trees. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 2336–2342, Portorož, Slovenia : European Language Resources Association (ELRA).

BRESNAN J., CUENI A., NIKITINA T. & BAAYEN R. (2004). Predicting the dative alternation. *Cognitive foundations of interpretation*.

CHAUDHARY A., ANASTASOPOULOS A., PRATAPA A., MORTENSEN D. R., SHEIKH Z., TSVETKOV Y. & NEUBIG G. (2020). Automatic extraction of rules governing morphological agreement. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 5212–5236, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.422.

CHAUDHARY A., SHEIKH Z., MORTENSEN D. R., ANASTASOPOULOS A. & NEUBIG G. (2022). Autolex : An automatic framework for linguistic exploration.

DE MARNEFFE M.-C., MANNING C. D., NIVRE J. & ZEMAN D. (2021). Universal dependencies. *Computational linguistics*, **47**(2), 255–308.

EFRON B., HASTIE T., JOHNSTONE I. & TIBSHIRANI R. (2004). Least angle regression. *Annals of Statistics*.

GERDES K., GUILLAUME B., KAHANE S. & PERRIER G. (2018). Sud or surface-syntactic universal dependencies : An annotation scheme near-isomorphic to ud. In *Universal Dependencies Workshop (UDW)*.

GUILLAUME B., DE MARNEFFE M.-C. & PERRIER G. (2019). Conversion et améliorations de corpus du français annotés en Universal Dependencies. *Revue TAL : traitement automatique des langues*, **60**(2), 71–95. HAL : hal-02267418.

MARKOWITZ H. (1952). Portfolio selection. *Journal of Finance*.

MEL'ČUK I. (1988). *Dependency Syntax : Theory and Practice*. State University of New York Press.

OSBORNE M. R., PRESNELL B. & TURLACH B. A. (2000). On the lasso and its dual. *Journal of Computational and Graphical statistics*, **9**(2), 319–337.

SHALEV-SHWARTZ S. & BEN-DAVID S. (2014). *Understanding machine learning : From theory to algorithms*. Cambridge university press.

THUILIER J. (2012). *Contraintes préférentielles et ordres des mots en français*. Thèse de doctorat, Paris 7. Dirigée par Laurence Danlos Laurence et Benoît Crabbé. Linguistique théorique, descriptive et automatique.