Traitement automatique des langues

Corpus annotés

sous la direction de
Marie Candito
Mark Liberman

Vol. 60 - n°2 / 2019

ATALA

Corpus annotés

ATALA

Revue de
l'Association
pour le Traitement
Automatique
des Langues

TAL
Vol.
60

n°2
2019

Corpus annotés

ATALA

Traitement automatique des langues

Revue publiée depuis 1960 par l'Association pour le Traitement Automatique des LAngues (ATALA), avec le concours du CNRS, de l'Université Paris VII et de l'Université de Provence

ISSN 1965-0906

http ://atala.org/revuetal

Traitement automatique des langues

Comité de rédaction

Secrétaire
Peggy Cellier - IRISA, INSA Rennes

Traitement automatique des langues

Volume 60 – n°2 / 2019

CORPUS ANNOTÉS

Table des matières

Introduction to the special issue on annotated corpora

Marie Candito* — **Mark Liberman****

* *LLF - Université Paris Diderot / CNRS*
** *University of Pennsylvania*

ABSTRACT. *Annotated corpora are increasingly important for linguistic scholarship, science and technology. This special issue briefly surveys the development of the field and points to challenges within the current framework of annotation using analytical categories as well as challenges to the framework itself. It presents three articles, one concerning the evaluation of the quality of annotation, and two concerning French treebanks, one dealing with the oldest project for French, the French Treebank, the second concerning the conversion of French corpora into the cross-lingual framework of Universal Dependencies, thus offering an illustration of the history of treebank development worldwide.*

RÉSUMÉ. *Les corpus annotés sont toujours plus cruciaux, aussi bien pour la recherche scientifique en linguistique que le traitement automatique des langues. Ce numéro spécial passe brièvement en revue l'évolution du domaine et souligne les défis à relever en restant dans le cadre actuel d'annotations utilisant des catégories analytiques, ainsi que ceux remettant en question le cadre lui-même. Il présente trois articles, l'un concernant l'évaluation de la qualité d'annotation, et deux concernant des corpus arborés du français, l'un traitant du plus ancien projet de corpus arboré du français, le French Treebank, le second concernant la conversion de corpus français dans le schéma interlingue des Universal Dependencies, offrant ainsi une illustration de l'histoire du développement des corpus arborés.*

KEYWORDS: *Annotated corpora, Resources for NLP, Linguistic resources*

MOTS-CLÉS : *Annotation de corpus, Ressources pour le TAL, Ressources linguistiques*

TAL. Volume 60 – n°2/2019, pages 7 à 17

1. From corpus-based theology, anthropology, and lexicography to modern NLP

Long before the invention of digital computers, the collection and annotation of corpora began as efforts to support the interpretation of culturally important texts, and continued as efforts to document language use. We'll introduce this special issue with a brief survey of some of this history, before providing some historical notes on annotated corpora for NLP in section 1.1, and for research in linguistics in section 1.2.

One well-known example of text augmented with interpretation comments is the Talmud, which, as Mielziner (1903) explains, "consists of two distinct works, the *Mishna*, as the text, and the *Gemara* as a voluminous collection of commentaries and discussions of that text". A different sort of religiously-motivated example is Strong (1890), which assigns a number to each of 8,674 Hebrew "root words" (what we would call "lemmas") used in the Old Testament, and similarly to each of 5,624 Greek lemmas in the New Testament, and annotates each relevant word of the English King James Version with the number of the source-language word. This allows someone to read the original texts, in some sense, even if they have little or no knowledge of the source languages. And because the whole thing is indexed as a concordance, they can find and compare all of the passages that use a given English, Hebrew, or Greek word.

An approach more similar in design to contemporary corpus annotation is the tradition of interlinear text, represented in a fully mature form in Müller *et al.* (1864). As that work's title states, it presents three layers of annotation for each word of the Sanskrit text of the *Hitopadeśa*: "transliteration, grammatical analysis, and English translation". As with the Talmud and Strong's concordance, Müller's work was based on a text that had existed for hundreds of years. In the late 19th and early 20th century, interlinear annotations of newly-collected texts became a standard tool in the development of linguistic anthropology and linguistic documentation. Franz Boas and his many followers saw the collection and annotation of texts as central to the understanding of languages and cultures (Boas and Hunt, 1905; Epps *et al.*, 2017).

Roberto Busa's 1946 dissertation (Busa, 1949) dealt with the concept of "presence" in the thought of Thomas Aquinas. Busa came to believe that he needed to understand the shades of meaning associated with Aquinas's use of the Latin preposition *in*, and based his dissertation on 10,000 examples of this usage, collected and written by hand on file cards. As he completed that work, Busa began to dream of a set of machine-readable cards a thousand times larger, from which automatic sorting could create unlimited opportunities for such corpus-based theology. In 1949, even before the first primitive digital computers became generally available, Busa persuaded IBM to support the creation of the *Index Thomisticus*, a complete lemmatized concordance to everything Aquinas wrote (Jones, 2016; Winter, 1999). The millions of punched cards constituting this corpus were completed in 1967 (Busa, 1974).

A similar transition from file cards to punched cards (and digital tape and onwards) took place in the field of lexicography. Starting in 1879, James Murray's "reading programme" turned mountains of books into thousands of "slips" documenting word usage, eventually organized into the *Oxford English Dictionary* (Winchester, 1998).

A similar program was carried out by the Merriam-Webster company in creating their series of dictionaries. In the early 1960s, Kučera and Francis created the Brown corpus (Kučera and Francis, 1967), a million-word balanced corpus of written American English meant to support "computer-based research in the English language"; and just a few years later, Houghton-Mifflin used the Brown corpus as the citation base for the first edition of the *American Heritage Dictionary* (Morris, 1969).

The idea of basing dictionaries and other forms of language documentation on large representative digital text corpora was soon adopted widely, since it was essentially a computational enhancement of methods that had long been in use. In Sweden, the pioneering work of Sture Allén at the University of Gothenburg in the 1960s led to Press-65, an electronic text corpus of one million words of newspaper text (Allén, 1968), and the establishment of Språkbanken (the Swedish Language Bank) in 1975. Also in the 1970s, researchers at Lund University created Talbanken, a syntactically-annotated corpus of Swedish (Einarsson, 1976a; Einarsson, 1976b). The Czech Academic Corpus project started in the 1970s at the Institute of the Czech Language [1]. The primary goal of this project was to create a corpus that would contain manual annotation of morphology and syntax of Czech, as a base for building a frequency dictionary. In the UK, the Collins Birmingham University International Language Database (COBUILD) project, begun in 1980 and funded by the Collins publishing company, formed the basis for the *Collins COBUILD Dictionary*, first published in 1987, and for a series of other reference works (Sinclair, 1987; Moon, 2009). And, beginning in 1991, Oxford University Press led a consortium of dictionary publishers, academic research centers, and government organizations in creating the British National Corpus (BNC), an open corpus of 100 million words from sources meant to be representative of spoken and written English in Great Britain in the last decade of the 20th century (Aston and Burnard, 1998). Part-of-speech and sense tagging were involved in the BNC project from the beginning (Atkins, 1993; Leech *et al.*, 1994; Kilgarriff, 1998).

In France, the *Trésor de la Langue Française* ("French language treasury", hereafter "TLF") project started in the 1960s to study French vocabulary usage, at the INALF laboratory in Nancy. As part of this project, the Frantext textual database of literary and technical texts was set up in order to provide examples for the TLF dictionary. After dictionary completion (the 16 volumes were released between 1971 and 1994, and are now an online resource, the "TLFi"), access was given to the textual database, first within the lab via a search engine in 1985, and later via online access in 1998 (according to Montémont [2008]). The current version contains over 5,000 automatically-parsed texts, totaling 250 million words [2]. And the MULTEXT project, funded by the European Commission in the early 1990s (Ide and Véronis, 1994), aimed to provide "generally usable software tools to manipulate and analyse text corpora and to create multilingual text corpora with structural and linguistic markup."

1. For a history of the project, see `http://ufal.mff.cuni.cz/rest/CAC/doc-cac10/cac-guide/eng/html/chapter2.html`.
2. `https://www.frantext.fr/`.

The MULTEXT plan included standards for encoding linguistic annotation, including morphology, syntax, parallel text alignment, and prosody.

1.1. *Corpora for Human Language Technology*

Turning to the use of corpora for human language technology, research on the statistics of messages, by engineers tasked with sending them efficiently turned to have a major impact – perhaps the most important one for the modern use of corpora in NLP, since it has resulted in a flowering of diverse approaches to the annotation of large linguistic datasets, used to train computer systems for tasks far beyond the problems of message encoding and transmission.

Shannon (1948) laid out a theory of optimal transmission over noisy channels, and showed that a simple empirical model could provide an arbitrarily close approximation to the relevant statistics of message sequences. Baum *et al.* (1970) and others provided computationally-efficient extensions of this model to cases where we can observe only a stochastic function of the hypothesized underlying text – and, over the course of the next few decades, such methods were applied to problems of speech recognition (Baker, 1975), machine translation (Brown *et al.*, 1990), part-of-speech tagging (Church, 1989), parsing (Lari and Young, 1990), and many other forms of speech and language analysis.

Starting in the mid-1980s, the US Defense Advanced Research Projects Agency (DARPA) began promoting research in this area, using what has been called the "Common Task method" (Liberman and Wayne, 2020) or "Shared Task method". This research management technique begins with a shared training dataset and an automated quantitative performance metric, with periodic competitive evaluations on test data withheld for the purpose. Because the datasets and evaluation software were published, and because the method succeeded in fostering gradual improvements in the targeted technologies, the Common Task approach has been widely adopted.

And large annotated corpora have been at the center of the process, since the dominant paradigm has been supervised machine learning, in which a system is trained and tested on a body of text or speech in which the desired analysis is explicitly presented (Liberman, 1991). Such analyses include morphosyntactic categories and relations in text, word senses, references to semantically-defined entities, textual evidence for adding information to a "knowledge graph", and many other things. The effectiveness of supervised models crucially depends on the availability of a large volume of annotated corpora. We focus below on the early days of two famous types of annotation, POS-annotated corpora and syntactic treebanks, whose development has been parallel to statistical taggers and parsers.

Perhaps the earliest example of digital corpus annotation was the part-of-speech (POS) tagging of the previously-mentioned Brown Corpus, which was motivated by language documentation and lexicography, and obtained by manually correcting the output of hand-written rules (Greene and Rubin, 1971). POS taggers based on statisti-

cal machine learning (Garside *et al.*, 1987; Church, 1989) were developed in the 1980s and early 1990s, as were broad-coverage statistically-trained parsers (Magerman and Marcus, 1990; De Marcken, 1990). These systems were trained on text collections provided with part-of-speech tags and syntactic analyses – and the same systems were also used in the creation of these treebanks to improve the productivity of human annotators by providing them with an automatically-created draft to correct. Morphosyntactically annotated corpora followed for a wide variety of typologically diverse languages.

The most well known and widely used syntactically-analyzed corpus has been the Penn Treebank (Marcus *et al.*, 1993), whose creation started in the late 1980s. The resulting dataset has been used for thousands of works on English syntax and systems for analyzing it – Google Scholar lists about 19,000 works that reference it. A crucial reason for this popularity (besides the economic importance of the English language) is that the Penn Treebank was made easily available to researchers all over the world, as the Brown corpus had been, in contrast to some other early collections such as the Swedish treebanks, which were tightly held by their creators. Similar treebanks were subsequently built for many typologically diverse languages.

A parallel strand of corpus creation has used syntactic dependencies (Tesnière, 1959) rather than syntactic constituents. From a formal point of view, these two representations are essentially equivalent, but the non-nested syntactic relations that are common in free-word-order languages are more easily represented by (crossing) dependencies than by (discontinuous) constituents. The Prague Dependency Treebank (with a first release in 1998 [Hajič, 1998]) was an important influence on the history of this approach, which has culminated in the Universal Dependencies project [3] (Nivre *et al.*, 2016), an attempt to provide cross-linguistically consistent dependency annotation. This open community effort with over 200 contributors has led to 146 treebanks covering 83 languages (as of version 2.4). Despite the inevitable approximations of the cross-lingual scheme, the resource is abundantly used for typological quantitative research and for cross-lingual parsing.

Over the past decade, so-called "deep learning" methods have achieved better performance than statistical machine-learning methods on most NLP tasks. But these methods are even hungrier for training data, and so the need for large annotated corpora to support Human Language Technology has only increased.

1.2. *Corpora for scientific and scholarly research*

The Child Language Data Exchange System (CHILDES), begun in 1984, established a repository for sharing records of child language acquisition (MacWhinney and Snow, 1985; MacWhinney, 2014), based on a system for discourse notation and coding called CHAT, and including transcripts of caregiver-child interactions expressed

3. `https://universaldependencies.org/`.

in that form, in some cases with audio and/or video recordings. The accumulation of shared material in CHILDES continues, with more than 130 different sources in 26 languages now available. The CHAT format allows for (but does not require) annotation of time codes, pronunciations, dysfluencies and speech errors, prosody, and speech act categories, among other things.

The study of historical syntax has always been corpus based, since historical texts provide the only concrete evidence of how language was used in earlier periods. Over the past twenty years, researchers at universities in the US, the UK, Finland, and Norway have created a series of parsed corpora totaling about 9 million words and covering more than a thousand years of linguistic history (Taylor, 2020). Similar efforts are underway for French, Portuguese, and Spanish.

Since the 1960s, research in quantitative sociolinguistics has been based on statistical modeling of annotation of the speech patterns of people varying in gender, age, location, socio-economic status, and communicative context. In earlier times, recordings and annotations were closely (and informally) held by the researchers that collected them, but, more recently, the culture of the discipline has begun changing in the direction of digital archiving and sharing of research datasets (Kendall, 2008). A similar trend has resulted in the digitization and (partial) publication of the archives of many dialect atlas projects from the past century (Nerbonne, 2009).

Thanks to the influence of psychology, as well as results within theoretical linguistics itself, such as Bresnan (2007), statistical information is increasingly viewed as part of the competence of speakers rather than a mere artefact of linguistic performance. The statistical exploration of corpora thus serves to validate linguistic hypotheses and even to discover new patterns. In such studies, sophisticated linguistic annotation may be needed in order to compile the needed counts.

2. Forward-looking perspective

Looking forward to the future of annotated corpora, we can first comment on some "sociological" aspects, both within the linguistics and NLP research communities. Annotation projects are costly and time-consuming, meaning reusability is crucial. NLP researchers tend to reuse already existing datasets, not only to enable comparison between systems but simply because it is easier. This might explain the emergence of annotation scheme standards, often stemming from English-centered projects, initiated by major American NLP players. Resources in other languages either use the same schemes, even if at a smaller scale, or an original annotation scheme, at the risk of lacking international visibility. Interestingly, the Universal Dependencies project, although initially driven by such players (Stanford, through the Stanford dependencies, and Google, through the universal POS tagset), has fostered a global reflection on an annotation scheme with a multilingual vocation.

From a practical point of view of easing linguistic resource production, there are currently several approaches, in particular using more efficient tools for experts, or us-

ing non-expert annotators. The possibility to leverage a potentially worldwide workforce via crowd-sourcing platforms has modified the economics of resource production. Note though that inequalities between languages are reinforced. In 2014, out of 100 languages, 13 only were considered to have an adequate workforce among the turkers (Pavlick *et al.*, 2014). Ethical concerns are now part of major NLP conferences and the focus of specific workshops or special issues (Fort *et al.*, 2016; Hovy *et al.*, 2017). *Games with a purpose* are an alternative used for varied linguistic resources including annotated corpora, for instance for coreference annotation (Chamberlain *et al.*, 2013) or dependency syntax (Guillaume *et al.*, 2016).

Another research lead is to make better use of the existing resources, in particular with more efficient learning from small datasets. Multilingual learning leverages data in several languages to better model the very same languages or some related ones.

Coping with multilinguality for resource production is indeed a major challenge, both practical and scientific. As already noted, the Universal Dependencies project succeeded in having hundreds of contributors collaborate. The PARSEME project has produced guidelines and data for 20 languages, for verbal multi-word expressions. Finally, annotating multimodal data is another challenge, with the necessity to annotate interconnections between different modalities (speech, gesture, emotional states...).

The framework of annotating corpora using analytic categories (for whatever kind of linguistic concept) is itself challenged, in particular given the current use of continuous representations in neural models. Frontiers between linguistic categories are often difficult to draw sharply (examples of well-known difficulties are the argument/adjunct distinction in syntax, or the adjective versus participle distinction in many languages, including French or German). This inevitably impacts the annotation process. For instance, Plank *et al.* (2014) show that disagreements on POS annotation concern debatable linguistic points rather than random errors, and should be used in the learning phase. These difficulties question the theory, but also the current annotation methodology, in which the resolution of annotation conflicts is the most time-consuming phase.

From the theoretical point of view, this indeterminacy of analytic category frontiers might be justified. Firstly, abandoning rigid frontiers between categories is empirically validated by the success of using continuous representations in neural NLP models, for all kinds of discrete symbols, enabling various mathematical representations of category combinations. Secondly, this can also lead to the idea of abandoning analytic categories altogether. The current trend in NLP is to overcome the need for annotated data, which is insufficient for most languages and tasks. Learning from raw texts and end-to-end models benefit from the enormous amount of raw texts, and challenge supervised NLP models trained on scarse but sophisticated symbolic annotations. But symbols do come back by the window: current research efforts on interpretability of neural models show that NLP without meta-linguistic explicitation is not satisfying. A promising research perspective is to integrate top-down (formalization to data) and bottom-up (from data to neural nets parameters) approaches.

3. Content of the special issue

The special issue contains three papers, two concerning syntactic treebanks and one concerning the evaluation of the quality of annotations. It is striking that the first treebank paper (Abeillé, Clément and Liégeois, *Un corpus arboré pour le français: le French Treebank*) provides an overview and some feedback on the first treebank project for French, namely the French Treebank, which has nourished NLP research for the French language, and whose development spans over the last twenty years. On top of an overview of the major linguistic choices underlying the annotation scheme, this paper is the occasion to provide some feedback on what could have been made differently, in the light of the various uses of the corpus in the last years. The paper presents the first full release of the corpus, namely the complete annotation of the whole corpus, namely meta-data concerning the author and domain of the articles, and, for each sentence, the annotation of multi-word expressions, parts-of-speech, morphological features, syntactic constituents and grammatical functions. Finally, a partial evaluation of the annotation quality is provided for the first time.

Interestingly, the other treebanking paper (Guillaume, de Marneffe and Perrier, *Conversion et amélioration de corpus du français annotés en* Universal Dependencies) focuses on how to have various treebanks converge within the multilingual annotation scheme of the *Universal Dependencies*. The difficulties of retaining linguistic description accuracy while using a scheme meant to be multilingual are described and illustrated with the French case. The paper presents a few annotation choices for cases not fully specified in the UD guidelines. A methodology for improving the quality of the resulting corpora is described, namely the double conversion method (converting into one annotation scheme, and converting back), thanks to the use of graph-rewriting rules. Differences with the original annotation signal errors either in the conversion rules, or in the original annotation.

The paper by Brégeon, Antoine, Villaneau and Halftmeyer, *Redonner du sens à l'accord inter-annotateur : vers une interprétation des mesures d'accord en termes de reproductibilité des annotations* focuses on annotation quality evaluation. Such an evaluation is essential to enable annotated corpora to serve as basis for valuable scientific findings. More precisely, the paper concerns the evaluation of a categorization task, in the multi-annotator setting, in which case the quality evaluation focuses on the inter-annotator agreement. The authors consider that popular chance-corrected measures such as Cohen's kappa and Krippendorff's alpha scores are difficult to interpret, and point out the arbitrary nature of the usual interpretation scale of the kappa. To get a more interpretable measure, the authors propose to evaluate the *stability* of the reference annotation. This is achieved by considering an average variation of the reference that would be obtained when taking a majority vote on subsets of annotators instead of on all annotators (note that this entails that at least three annotators per item are required in order to take subsets).

Experiments conducted both on real and simulated data show a correlation between the multi-annotator kappa score and the proposed reproductibility score, which

the authors consider more interpretable. On a closer look though, the proposed metric varies, for the same kappa value, according to the distribution of divergences, i.e. according to whether the divergences are concentrated on a few items or scattered on many of them. On the former case, the variation rate tends to augment. So, despite its sensitivity to the number of classes and number of annotators to consider in subsets, the metric is proved to provide additional information with respect to the kappa.

Acknowledgements

We are very grateful to the editorial board and to the reviewing committee of the *TAL* journal, with special thanks to Emmanuel Morin.

We would also like to warmly thank the members of this issue's specific reviewing committee: Pascal Amsili (LLF, Université Paris Diderot), Farah Benamara (IRIT, Université Toulouse III - Paul Sabatier), Christophe Benzitoun (ATILF, Université de Lorraine), Delphine Bernhard (LiLPa, Université de Strasbourg), Kim Gerdes (ILPGA, Université Sorbonne Nouvelle), Marie-Catherine de Marneffe (Ohio State University), Paola Merlo (Université de Genève), Thomas François (Université catholique de Louvain), Carlos Ramisch (LIS, Aix-Marseille Université), Benoît Sagot (Almanach, INRIA), Agata Savary (LIFAT, Université de Tours), Djamé Seddah (Almanach, INRIA), Marie Tahon (LIUM, Université du Mans).

4. References

Allén S., "Report on work in computational linguistics at the University of Göteborg", *in* E. Mater, J. Štindlová (eds), *Les Machines dans la linguistique*, Éditions de l'Académie Tchécoslovaque des Sciences, Prague, 1968.

Aston G., Burnard L., *The BNC handbook: exploring the British National Corpus with SARA*, Edinburgh University Press, 1998.

Atkins S., "Tools for computer-aided corpus lexicography: the Hector project", *Acta Linguistica Hungarica*, vol. 41, n° 1-4, p. 5, 1993.

Baker J. K., Stochastic modeling as a means of automatic speech recognition, PhD thesis, Carnegie-Mellon University, 1975.

Baum L. E., Petrie T., Soules G., Weiss N., "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains", *The Annals of mathematical statistics*, vol. 41, n° 1, p. 164-171, 1970.

Boas F., Hunt G., *Kwakiutl texts*, vol. 5, EJ Brill, 1905.

Bresnan J., "Is syntactic knowledge probabilistic?", *in* S. Featherston, W. Sternefeld (eds), *Roots: Linguistics in Search of Its Evidential Base*, Mouton de Gruyter, 2007.

Brown P. F., Cocke J., Della Pietra S. A., Della Pietra V. J., Jelinek F., Lafferty J. D., Mercer R. L., Roossin P. S., "A statistical approach to machine translation", *Computational linguistics*, vol. 16, n° 2, p. 79-85, 1990.

Busa R., *La Terminologia tomistica dell'interiorità (Saggi di metodo per un'interpretazione della metafisica della presenza)*, Archivum Philosophicum Aloisianum, 1949.

Busa R., "Index Thomisticus Sancti Thomae Aquinatis Operum Omnium Indices Et Concordantiae in Quibus Verborum Omnium Et Singulorum Formae Et Lemmata Cum Suis Frequentiis Et Contextibus Variis Modis Referuntur", 1974.

Chamberlain J., Fort K., Kruschwitz U., Lafourcade M., Poesio M., "Using Games to Create Language Resources", *in* Gurevych, Iryna, Kim, Jungi (eds), *Theory and Applications of Natural Language Processing*, Springer, 2013.

Church K. W., "A stochastic parts program and noun phrase parser for unrestricted text", *International Conference on Acoustics, Speech, and Signal Processing*, p. 695-698, 1989.

De Marcken C. G., "Parsing the LOB corpus", *ACL*, p. 243-251, 1990.

Einarsson J., Talbankens skriftspråkskonkordans, Technical report, Lund University: Department of Scandinavian Languages, 1976a.

Einarsson J., Talbankens talspråkskonkordans, Technical report, Lund University: Department of Scandinavian Languages, 1976b.

Epps P. L., Webster A. K., Woodbury A. C., "A holistic humanities of speaking: Franz Boas and the continuing centrality of texts", *International Journal of American Linguistics*, vol. 83, n° 1, p. 41-78, 2017.

Fort K., Adda G., Bretonnel Cohen K., "Éthique et traitement automatique des langues et de la parole : entre truismes et tabous", *Traitement Automatique des Langues*, vol. 57, n° 2, p. 7-19, 2016.

Garside R., Leech G., Sampson G., *The Computational Analysis of English: A Corpus-Based Approach*, Longman, 1987.

Greene B., Rubin G., *Automatic Grammatical Tagging of English*, Department of Linguistics, Brown University, 1971.

Guillaume B., Fort K., Lefèbvre N., "Crowdsourcing Complex Language Resources", *COLING*, p. 3041-3052, 2016.

Hovy D., Spruit S., Mitchell M., Bender E. M., Strube M., Wallach H., "Proceedings of the First ACL Workshop on Ethics in Natural Language Processing", Valencia, 2017.

Ide N., Véronis J., "MULTEXT: Multilingual text tools and corpora", *COLING*, p. 588-592, 1994.

Jones S. E., *Roberto Busa, SJ, and the emergence of humanities computing: the priest and the punched cards*, Routledge, 2016.

Kendall T., "On the history and future of sociolinguistic data", *Language and Linguistics Compass*, vol. 2, n° 2, p. 332-351, 2008.

Kilgarriff A., "Gold standard datasets for evaluating word sense disambiguation programs", *Computer Speech & Language*, vol. 12, n° 4, p. 453-472, 1998.

Kučera H., Francis W. N., *Computational analysis of present-day American English*, Brown University Press, 1967.

Lari K., Young S. J., "The estimation of stochastic context-free grammars using the inside-outside algorithm", *Computer speech & language*, vol. 4, n° 1, p. 35-56, 1990.

Leech G., Garside R., Bryant M., "CLAWS4: the tagging of the British National Corpus", *The 15th International Conference on Computational Linguistics (COLING 1994)*, 1994.

Liberman M., Wayne C., "Human Language Technology", *AI Magazine*, 2020.

Liberman M. Y., “The trend towards statistical models in natural language processing”, *Natural Language and Speech*, Springer, p. 1-7, 1991.

MacWhinney B., *The CHILDES project: Tools for analyzing talk, Volume II: The database*, Psychology Press, 2014.

MacWhinney B., Snow C., “The child language data exchange system”, *Journal of child language*, vol. 12, nº 2, p. 271-295, 1985.

Magerman D. M., Marcus M. P., “Parsing a Natural Language Using Mutual Information Statistics.”, *AAAI*, vol. 90, p. 984-989, 1990.

Marcus M. P., Marcinkiewicz M. A., Santorini B., “Building a Large Annotated Corpus of English: The Penn Treebank”, *Computational Linguistics*, 1993.

Mielziner M., *Introduction to the Talmud*, Funk & Wagnalls, 1903.

Moon R., *Words, grammar, text: revisiting the work of John Sinclair*, vol. 18, John Benjamins Publishing, 2009.

Morris W. (ed.), *The American Heritage Dictionary of the English Language*, Houghton-Mifflin, 1969.

Müller F. M. *et al.*, *The First Book of the Hitopadeśa containing the Sanskrit text, with interlinear transliteration, grammatical analysis, and English translation [edited by Max Müller]*, vol. 1, Longman, Green, Longman, Roberts & Green, 1864.

Nerbonne J., “Data-driven dialectology”, *Language and Linguistics Compass*, vol. 3, nº 1, p. 175-198, 2009.

Nivre J., de Marneffe M.-C., Ginter F., Goldberg Y., Hajič J., Manning C. D., McDonald R., Petrov S., Pyysalo S., Silveira N., Tsarfaty R., Zeman D., “Universal Dependencies v1: A Multilingual Treebank Collection”, *LREC*, 2016.

Pavlick E., Post M., Irvine A., Kachaev D., Callison-Burch C., “The Language Demographics of Amazon Mechanical Turk”, *Transactions of the Association for Computational Linguistics*, vol. 2, p. 79-92, 2014.

Plank B., Hovy D., Søgaard A., “Linguistically debatable or just plain wrong?”, *ACL*, p. 507-511, 2014.

Shannon C. E., “A mathematical theory of communication”, *Bell system technical journal*, vol. 27, nº 3, p. 379-423, 1948.

Sinclair J. M., *Looking up: An account of the COBUILD project in lexical computing and the development of the Collins COBUILD English language dictionary*, Collins Elt, 1987.

Strong J., *The Exhaustive Concordance of the Bible*, Hodder and Stoughton, 1890.

Taylor A., “Treebanks in Historical Syntax”, *Annual Review of Linguistics*, 2020.

Tesnière L., *Éléments de syntaxe structurale*, Klincksieck, Paris, 1959.

Winchester S., *The Professor and the Madman*, Harper, 1998.

Winter T. N., “Roberto Busa, SJ, and the invention of the machine-generated concordance”, *Faculty Publications, Classics and Religious Studies Department,* p. 70, 1999.

Un corpus arboré pour le français : le French Treebank

Anne Abeillé* — **Lionel Clément**** — **Loïc Liégeois*****

* *Laboratoire de Linguistique Formelle (LLF), Université Paris Diderot*
anne.abeille@univ-paris-diderot.fr
** *LaBRI , Université Bordeaux*
lionel.clement@u-bordeaux.fr
*** *CLILLAC-ARP et LLF, Université Paris Diderot*
loic.liegeois@univ-paris-diderot.fr

RÉSUMÉ. *Nous présentons un bilan du Corpus arboré du français, ou French Treebank (FTB) (1996-2016), qui est une ressource lexicale et syntaxique unique en son genre, richement annotée (et validée manuellement) pour les linguistes, et pour le TAL, avec environ 300 utilisateurs dans le monde. Après avoir exposé les principes de construction, et les principaux choix d'annotation, nous présentons l'état final du corpus, ses différents formats, et une première évaluation. Nous présentons aussi quelques ressources dérivées et des exemples d'interrogation.*

ABSTRACT. *We present a review of the French Treebank (FTB) (1996-2016), a lexical and syntactic resource with rich annotation and manual validation, which is usable by linguists and for NLP and has about 300 users in the world. We summarize the building principles and the main annotation choices, and describe the final version, the different formats and a first evaluation. We also present some derived resources and some query examples.*

MOTS-CLÉS : *corpus arboré, français, syntaxe.*

KEYWORDS: *treebank, French, syntax.*

1. Introduction

Nous présentons la version finale du French Treebank (FTB), projet initié à l'université Paris Diderot en 1996, avec le soutien de l'IUF, du CNRS, du LLF, de la DGLFLF et du CNRTL, et achevé en 2016.
Le projet a joué un rôle pionnier dans l'« outillage » de la langue française (Habert, 2004), car aucun corpus de référence n'existait à l'époque pour la syntaxe du français. Il consistait à annoter des textes écrits, sur le modèle de l'annotation du *Wall Street Journal* effectuée dans le cadre du Penn Treebank (Taylor *et al.*, 2003). Le journal *Le Monde* a été choisi pour sa disponibilité, son écriture soignée avec très peu de fautes typographiques, et la variété des sujets abordés.
Après avoir détaillé les différentes couches d'annotation (mots, syntagmes, fonctions), nous présentons la version finale du corpus avec ses différents formats, y compris la version en dépendances, et quelques chiffres concernant la distribution des catégories et des syntagmes. Dans un second temps, nous présentons une première évaluation, pour les différents niveaux d'annotation, ainsi que les principales utilisations du corpus, y compris les ressources dérivées, et enfin une comparaison avec d'autres corpus français plus récents annotés pour la syntaxe.

2. Les différentes couches d'annotation du FTB

Dans un premier temps ont été réalisées les annotations lexicales (catégories, sous-catégories, flexion, mots composés avec composants) (Abeillé et Clément, 1999b) ; puis les annotations syntaxiques (constituants majeurs, fonctions grammaticales) (Abeillé *et al.*, 2000 ; Abeillé *et al.*, 2003 ; Abeillé et Barrier, 2004). Elles ont été annotées par des outils dédiés (*taggeur*, *chunker*, étiqueteur fonctionnel) (Kinyon, 2001 ; Toussenel, 2001 ; Clément, 2001) et validées à la main, avec double correction par des annotateurs linguistes. Dans un dernier temps, ont été ajoutées des métadonnées pour chacun des articles : auteur, date, domaine.

2.1. *L'annotation lexicale*

L'annotation lexicale est particulièrement riche puisqu'elle comporte non seulement la catégorie (ou *part-of-speech*, POS) mais aussi la sous-catégorie, le lemme, la flexion et les catégories internes aux mots composés.
L'annotation automatique a été réalisée par des outils dédiés à cet effet, puis validée par des annotateurs linguistes. Une première étape a concerné la segmentation en phrases et en mots, avec validation manuelle pour tous les mots composés. Pour l'annotation des catégories, des sous-catégories et de la flexion, a été utilisé un *taggeur* dédié (Clément, 2001), fondé sur (Brill, 1993), avec plus 322 règles contextuelles écrites à la main, et un jeu réduit de 103 étiquettes, avec un taux d'erreur estimé à l'époque à 8 %. Les annotations ont été validées et enrichies par des annotateurs linguistes, puis complétées pour les lemmes avec des dictionnaires externes. Les composants de com-

posés (avec un jeu d'étiquettes réduit) ont fait l'objet d'une campagne d'annotation spécifique. Les outils informatiques et les procédures de validation sont décrits dans (Clément, 2001 ; Abeillé *et al.*, 2003) et les consignes d'annotation dans les guides associés au corpus (Abeillé et Clément, 1999a). Nous présentons ici les principaux choix d'annotation.

2.1.1. *L'annotation des catégories lexicales*

Le corpus compte 11 catégories lexicales, auxquelles s'ajoutent les étiquettes ET (mot étranger) et PONCT (ponctuation). La plupart des catégories, sauf Interjection, Verbe et Préposition, comportent des sous-catégories, par exemple subordination ou coordination pour Conjonction et propre ou commun pour Nom. L'ensemble des catégories avec leurs sous-catégories donne 41 étiquettes, et 218 une fois ajoutées les informations flexionnelles (voir tableau1). Par comparaison, un corpus comme le Penn Treebank comporte 36 étiquettes, en raison de la morphologie moins riche de l'anglais et d'un plus faible nombre de sous-catégories. À titre d'exemple pour le français, un corpus comme Frantext catégorisé comporte 22 étiquettes (voir section 6.1).

Ont été distinguées deux catégories : clitique (CL) pour les pronoms faibles, et PRO pour les pronoms forts. L'annotation morphologique prend en compte la flexion mais non la dérivation. L'étiquette PRE (pour préfixe) n'est utilisée que pour des préfixes détachables, écrits avec un trait d'union, parfois issus de mots comme *franco-*, pour *franco-allemand*.
Les mots étrangers intégrés à la syntaxe de la phrase sont étiquetés comme des mots français. Ceux qui ont l'étiquette ET sont ceux pour lesquels aucune catégorie ne peut être restituée, par exemple parce qu'ils sont en citation (*Errare humanum est* par exemple).

2.1.2. *L'annotation de la flexion*

Les informations de genre et de nombre sont ajoutées aux Adjectifs, Déterminants, Noms et Pronoms, plus la personne pour les possessifs et les Pronoms. Outre le nombre et la personne, les formes verbales sont aussi annotées pour le mode et le temps, ainsi que pour le genre pour les participes passés. Nous notons ces informations même si les formes ne sont pas distinctes. Ainsi, en contexte, la forme *rouge* sera notée comme masculin ou féminin, la forme *peux* comme 1re ou 2e personne du singulier, etc. Dans le cas des Pronoms, l'annotation se fait selon leur antécédent (*qui* relatif reçoit ainsi les mêmes genre, nombre et personne que son antécédent) ou leur référent (*je* reçoit l'information de genre correspondant à celui du locuteur). Dans le cas des noms propres, par exemple de marque ou de ville, l'information de genre n'est pas toujours disponible et reste alors non renseignée (Maurel et Belleil, 1996). Avec le recul, il aurait été plus facile de regrouper les numéraux sous une seule étiquette au lieu de quatre (Det, A, N, PRO). Il aurait aussi été intéressant de distinguer le *il* ou *ce* impersonnel ainsi que participe passé et participe passif, ce qui a été fait dans des versions ultérieures (Ribeyre *et al.*, 2014).

Catégorie	Étiquette	Sous-catégories	Flexion	Exemples
Adjectif	A	card, excl, indéf, inter, ord, poss, qual	genre, nomb, pers	*facile, mien, quelques, trois, troisième*
Adverbe	ADV	-, excl, inter, nég	-	*bien, heureusement, si*
Clitique	CL	objet, sujet, réfl	genre, nomb, pers	*je, toi, se*
Conjonction	C	coord, sub	-	*et, mais, que, si*
Déterminant	D	card, dém, déf, excl, indéf, inter, part, poss, nég	genre, nomb, pers	*ces, la, un, quel*
Mot étranger	ET	-	-	*and, uno*
Interjection	I	-	-	*hélas*
Nom	N	commun, propre	genre, nomb	*France, prix*
Pronom	PRO	card, dém, indéf, inter, nég, pers, poss, rel	genre, nomb, pers	*trois, tout, lequel, rien, eux*
Ponctuation	PONCT	fort, faible	-	. ! ? ,
Préfixe	PREF	-	-	*franco-, outre-*
Préposition	P	-	-	*à, de, sur*
Verbe	V	-	genre, nomb, pers, mode, temps	*avoir, montrait, pourra*

Tableau 1. *Les catégories morphosyntaxiques du FTB*

2.1.3. *L'annotation des lemmes*

Une fois les étiquettes morphosyntaxiques validées, les lemmes sont ajoutés automatiquement, avec un dictionnaire externe, et les rares cas restant ambigus (*suis* du verbe *suivre* ou du verbe *être*, *étaient* du verbe *étayer* ou du verbe *être*, par exemple) ont été résolus à la main.

2.1.4. *L'annotation des mots composés*

La segmentation en mots (tokens) considère tous les signes de ponctuation et les espaces comme des séparateurs, mais ne sépare pas *au* ou *du*. Ensuite, certains mots composés (*aujourd'hui, pomme de terre*) sont annotés comme composants et regroupés au sein d'un mot composé (« *compound* »). L'annotation des mots composés inclut les locutions et les « mots agglomérés » (Fradin, 2003), selon un ensemble de critères graphiques, morphologiques, syntaxiques et sémantiques définis dans le guide d'annotation (Abeillé et Clément, 1999a). La première phase d'annotation, automatique, a

été réalisée à l'aide des dictionnaires externes du LADL (Silberztein, 1993), puis enrichie et validée à la main. En effet, une séquence candidate n'est pas toujours un mot composé, comme pour *bien que* dans une phrase telle que *Juppé voudrait **bien que** quelqu'un l'aime.* Dans la version finale, les composants de composés ont tous une catégorie interne (catint) ; en revanche, ils n'ont ni sous-catégorie, ni lemme associé (voir figure 1). Ce niveau d'annotation permet d'étudier en tant que telle la formation des mots composés, mais aussi de dériver des versions du corpus en ignorant les mots composés (Schluter et van Genabith, 2007), ou en ne retenant que les mots composés grammaticaux et/ou irréguliers (Candito *et al.*, 2010).
Les mots composés reçoivent la même richesse d'annotation que les mots simples (POS, flexion, lemme) mais aussi une étiquette spécifique « *compound* » . En tant que mot composé, *bien que* reçoit l'étiquette CS et deux étiquettes internes (catint) : Adv pour *bien* et CS pour *que*.

Les discontinuités éventuelles sont notées avec les balises <next> et <prev>, comme pour *à cause, notamment, de*, où la préposition *à cause de* est coupée par un adverbe. Au total, le corpus compte 59 mots composés discontinus.

Le corpus comporte des mots composés grammaticaux (*peut-être, bien que*) mais aussi ce qu'on appelle aujourd'hui des entités nommées comme (*Parti socialiste*) (Sagot *et al.*, 2012). Certains sont très longs (par exemple *Direction de la consommation, de la concurrence et de la répression des fraudes* ou *Société des autoroutes du nord et de l'est de la France*).
Dans sa version actuelle, le corpus compte 32 546 mots composés, c'est-à-dire près de 1,5 mot composé par phrase. Avec le recul, certains noms composés auraient pu être considérés comme des combinaisons de mots simples, parce qu'ils respectent la syntaxe ordinaire, et certaines locutions verbales auraient pu être décomposées également. De plus, le critère de figement du nom pour les locutions verbales (*rendre compte* est figé mais pas *tirer (un bon) parti*) n'a pas toujours été bien suivi par les annotateurs, d'où certaines incohérences.

2.2. *L'annotation syntaxique*

L'annotation syntaxique comporte le découpage en constituants (syntagmes) et les principales fonctions grammaticales. L'annotation automatique a été réalisée par des outils dédiés à cet effet, puis validée par des annotateurs linguistes. Une première étape a concerné le découpage en syntagmes, avec validation manuelle pour les frontières de syntagmes et pour leur catégorie. L'étiqueteur syntaxique (*chunker*) fondé sur (Kinyon, 2001) identifiait les constituants majeurs (12 étiquettes), sans récursion. Évalué sur 500 phrases corrigées tirées au hasard, il avait 60 % de précision, 92,9 % de rappel et 94 % d'étiquettes correctes pour les bornes ouvrantes, et 60 % de précision, 58 % de rappel, 56.5 % d'étiquettes correctes pour les bornes fermantes (Toussenel, 2001 ; Clément, 2001 ; Abeillé *et al.*, 2003). L'étiqueteur fonctionnel (8 étiquettes) s'appuyait sur les syntagmes et 115 règles avec unification, écrites à la main. Il a été évalué sur un échantillon de 1 000 phrases corrigées, avec une préci-

sion moyenne de 89,69 % (max 99,47 % pour Sujet) et un rappel moyen de 89,27 % (max 95,48 % pour Modifieur) (Abeillé et Barrier, 2004).
Les consignes d'annotation sont présentées en détail dans les guides (Abeillé *et al.*, 1999 ; Abeillé, 2004). Nous présentons ici les principaux choix d'annotation, qui visaient à être compatibles avec plusieurs théories syntaxiques, de façon à ce que le corpus soit aisément convertible en différentes versions.

Catégorie	Étiquette	Fonctions	Exemples
Syntagme adjectival	AP	ATS, ATO, OBJ, MOD	*fier de lui, très grand*
Syntagme adverbial	AdP	OBJ, P-OBJ, MOD	*très bien, plus vite*
Syntagme coordonnant	COORD	SUJ, OBJ, A-OBJ, DE-OBJ, P-OBJ, MOD	*et la France, ni le Maroc ni la Tunisie*
Syntagme nominal	NP	ATS, ATO, SUJ, OBJ, A-OBJ, DE-OBJ, P-OBJ, MOD	*la France, plus de 30 %, tout cela*
Syntagme prépositionnel	PP	ATS, ATO, A-OBJ, DE-OBJ, P-OBJ, MOD	*à midi, en France*
Subordonnée	Sint, Srel, Ssub	ATS, ATO, SUJ, OBJ, A-OBJ, DE-OBJ, P-OBJ, MOD	*dont on parle, dit-on, quand il faudra*
Phrase racine	SENT	-	*Rien n'avance.*
Noyau verbal	VN	SUJ, OBJ, A-OBJ, DE-OBJ, P-OBJ, SUJ/OBJ, SUJ/A-OBJ, etc.	*j'ai vu, on parle, en avoir*
Syntagme verbal	VPinf, VPpart	ATS, ATO, SUJ, OBJ, A-OBJ, DE-OBJ, P-OBJ, MOD	*tout finir, de voir cela, en avançant*

Tableau 2. *L'annotation syntaxique du FTB*

2.2.1. *L'annotation des syntagmes*

Le choix a été fait de structures syntaxiques relativement plates. Au sein du syntagme nominal, le déterminant, le nom et ses dépendants sont au même niveau. Au sein de la phrase, sujet et compléments sont au même niveau également. La plupart des mots des catégories ouvertes (Adjectif, Nom, Verbe) projettent un syntagme même quand ils sont employés seuls : un nom propre ou un nom attribut correspond à un NP, un verbe intransitif à un VN (exemple 1a). Un adjectif attribut ou épithète postnominal correspond à un AP (exemple 1b), mais pas un épithète prénominal (exemple 1c) car il ne peut pas apparaître avec un dépendant avant le nom : *une facile *(à remporter) victoire* vs *une victoire facile (à remporter)* (Abeillé et Godard, 1999). Enfin, un adverbe ne correspond pas à un AdP quand il est seul, car les adverbes ont une distribution différente de celle des syntagmes adverbiaux (Abeillé et Godard, 2001). Un pronom faible (Clitique) appartient au VN, contrairement à un pronom fort (*lui*) qui projette un NP (exemple 1c) (Miller, 1992).

(1) a. [Paul $_{NP}$] [dort $_{VN}$] bien [dans [sa chambre $_{NP}$] $_{PP}$].
b. [La France $_{NP}$] [est $_{VN}$] [riche $_{AP}$].
c. [On a eu $_{VN}$] [un autre problème [important $_{AP}$] $_{NP}$] [avec [lui $_{NP}$] $_{PP}$].

La catégorie Syntagme verbal (VP) est réservée aux syntagmes subordonnés à l'infinitif (VPinf) ou au participe (VPpart). Les verbes conjugués projettent un noyau verbal (VN), qui comprend les auxiliaires, les participes et les clitiques, mais ne projettent pas de syntagme verbal (exemple 1c).

Le corpus ne comporte pas de catégories vides : les infinitifs n'ont pas de sujet implicite annoté, ni les impératifs. Nous avons une catégorie COORD pour les syntagmes coordonnés. Ainsi, ils peuvent être inclus dans un autre syntagme, ou être détachés (exemple 2c) (Abeillé, 2005). Il en résulte une structure symétrique pour les coordinations redoublées (exemple 2b), et asymétrique pour les autres (exemple 2a) (Mouret, 2007 ; Mouret, 2005).

(2) a. [le Maroc [et [la Tunisie $_{NP}$] $_{COORD}$] $_{NP}$]
b. [et [le Maroc $_{NP}$] $_{COORD}$] [et [la Tunisie $_{NP}$] $_{COORD}$]
c. [Il est parti $_{VN}$], [et vite $_{COORD}$].

Les frontières de syntagmes indiquent les enchâssements. Ainsi, un syntagme prépositionnel inclut généralement un syntagme nominal et un syntagme verbal un noyau verbal. Selon que le syntagme prépositionnel est inclus dans un syntagme nominal ou non, on distingue complément de verbe (exemple 3a) et complément de nom (exemple 3b). De même l'inclusion d'une subordonnée relative (Srel) dans un syntagme nominal indique son rattachement. Lorsque les deux analyses sont possibles, par exemple dans une construction à verbe support (Gross, 1976), l'annotation la plus plate a été retenue (exemple 3c).

(3) a. [Il y a $_{VN}$] [30 élèves $_{NP}$] [dans [cette classe $_{NP}$] $_{PP}$].
b. [On a arrêté $_{VN}$] [six [d'entre [eux $_{NP}$] $_{PP}$] $_{NP}$].
c. [Ce pays $_{NP}$] [a commis $_{VN}$] [des agressions $_{NP}$] [contre [ses voisins $_{NP}$] $_{PP}$].

2.2.2. *L'annotation des fonctions syntaxiques*

Les fonctions syntaxiques des syntagmes dépendant de verbes ont été annotées avec un outil dédié (Abeillé et Barrier, 2004) et corrigées à la main. Les syntagmes dépendant d'autres catégories ne portent pas de fonction. Celle-ci peut se déduire en partie du découpage en constituants : un syntagme nominal inclus dans un syntagme prépositionnel est complément de cette préposition ; une relative est modifieur du nom du syntagme nominal qui l'inclut, un adjectif aussi. Pour les syntagmes prépositionnels, la distinction entre modifieur et complément est toujours difficile ; elle a été faite

pour les dépendants de verbe (exemple 3a), pour lesquels existent des tests (obligatoire ou non, mobile ou non, remplaçable par un clitique ou non, etc.), mais non pour les dépendants de nom ou d'adjectif. Parmi les compléments prépositionnels de verbe, nous distinguons ceux en *à* (fonction A-OBJ), ceux en *de* (fonction DE-OBJ) et les autres (fonction P-OBJ). Un complément de lieu est noté P-OBJ, car la préposition pourrait être remplacée par une autre (exemple 2a). Il en résulte un jeu réduit de 8 étiquettes fonctionnelles, auxquelles s'ajoutent des combinaisons de fonctions dans le cas des séquences de clitiques (voir tableau 2), et 12 autres fonctions dans la version convertie en arbres de dépendance (voir section 3.1.4).
Quand un mot seul ne projette pas un syntagme, il ne reçoit pas de fonction. Avec le recul, le fait qu'un adverbe seul ne projette pas de syntagme empêche de distinguer adverbe complément (*aller bien*) et modifieur (*travailler bien*). Les clitiques sont inclus dans le noyau verbal, qui porte leur fonction. C'est le seul type de syntagme à pouvoir porter des fonctions combinées, par exemple Sujet et Objet (SUJ/OBJ), s'il contient plusieurs clitiques (exemple 4a). Les clitiques figés, qui ne correspondent pas à un complément, comme le *y* de *il y a*, ou les réfléchis intrinsèques ne portent pas de fonction. Le *il* impersonnel, en revanche, correspond à la fonction Sujet.

(4) a. [On leur a expliqué $_{\text{VN :SUJ/A\_OBJ}}$] [en détail $_{\text{PP :MOD}}$].
b. [Il y a $_{\text{VN :SUJ}}$] [des problèmes [de courant] $_{\text{NP :OBJ}}$] [à Moscou $_{\text{PP :P-OBJ}}$].
c. [Quelles solutions $_{\text{NP :OBJ}}$] [la France $_{\text{NP :SUJ}}$] [peut-elle $_{\text{VN :SUJ}}$] [proposer $_{\text{VP :OBJ}}$] ?
d. [La France $_{\text{NP :SUJ}}$] [en a perdu $_{\text{VN :OBJ}}$] [plusieurs $_{\text{NP :OBJ}}$].

Un verbe peut avoir deux sujets ou deux objets. Dans le cas de l'inversion complexe, le sujet nominal porte la fonction SUJ, tout comme le clitique inclus dans le noyau verbal (exemple 4c). Il n'y a pas de syntagmes discontinus et la même fonction Objet est annotée pour le clitique *en* et le pronom postverbal (exemple 4d). Les dépendances à distance ne sont pas notées non plus (Candito et Seddah, 2012a). Ainsi, dans l'exemple 4c, le syntagme initial est noté OBJ, mais sans préciser qu'il s'agit d'un complément de *proposer* et non de *peut-elle*.

2.3. *L'ajout des métadonnées*

En 2016, le FTB a été enrichi de métadonnées indiquant la date de parution, l'auteur et le domaine de chacun des articles, qui n'étaient pas disponibles à l'origine. Le corpus est constitué de 1 143 extraits d'articles du journal *Le Monde* pris aléatoirement dans les versions distribuées à l'époque, entre janvier 1990 et août 1993 : la moitié en 1992, un quart en 1990 et un quart en 1993. Cette période courte permet une bonne homogénéité et limite les variations diachroniques éventuelles. Chaque extrait contient en moyenne près de 19 phrases, 6 ne sont composés que d'une seule phrase tandis que l'extrait le plus long en comporte 28.

Un peu plus de la moitié des articles (579) ne sont pas signés et l'auteur est renseigné comme « LeMonde ». Les 564 autres ont 210 auteurs ou groupe d'auteurs différents. Si la majorité des auteurs (134) n'ont écrit qu'un seul article, certains sont surreprésentés, comme F. Renard (27 articles), A. Lebaude (23 articles) ou M. Colonna d'Istria (19 articles). L'annotation fondée sur le codage interne du *Monde* répartit les textes dans 14 domaines différents (Illouz *et al.*, 2000). Près de 80 % (912) des articles traitent de problématiques économiques : 737 viennent des pages « Économie » et 175 du supplément *Le Monde Économie*. Puis vient la politique étrangère (77 extraits) et le supplément *Le Monde Initiatives* (36 extraits).
Le corpus arboré French Treebank apparaît donc comme un corpus homogène : genre discursif, sources des articles de presse, années d'écriture des textes, etc. Avec le recul, le choix des articles aurait pu être plus équilibré.

3. La version 1.0 du FTB

Le Laboratoire de Linguistique Formelle (UMR 7110) a publié, en décembre 2016, la version finale du corpus « French TreeBank 1.0 », disponible en différents formats : XML d'origine, TIGER-XML, Penn TreeBank et CoNLL pour la version en dépendances (Candito *et al.*, 2009). Un site dédié permet de télécharger la ressource, recense la documentation disponible, et comporte une plateforme d'interrogation pour des requêtes lexicales et/ou syntaxiques (`ftb.linguist.univ-paris-diderot.fr`).
Le corpus est distribué gratuitement à des fins de recherche et avec une licence payante à des fins commerciales. Il compte plus de 300 utilisateurs dans le monde, dont la moitié sont inscrits depuis décembre 2016, avec en moyenne 5,6 nouveaux utilisateurs par mois.
Le corpus regroupe 21 550 phrases pour 644 595 tokens au total, signes de ponctuation compris, et 557 518 tokens, signes de ponctuation exclus. Ces chiffres sont calculés en comptant chaque composant de composé comme 1 token (par exemple, *la plupart* compte pour 2 tokens) et chaque amalgame (*au, du*) pour 2 également (*à* et *le* par exemple). Comme dans d'autres corpus journalistiques, la longueur moyenne des phrases est élevée, près de 26 mots par phrase.

3.1. *Les différents formats du FTB*

Afin de permettre son utilisation à la fois pour des études linguistiques et pour des tâches de TAL, le FTB a été structuré dans plusieurs formats standard. Le format XML d'origine a ainsi été progressivement converti aux formats TIGER-XML, Penn TreeBank et CoNLL.

3.1.1. *Le format XML d'origine*

Le corpus FTB a été au départ structuré au format XML (*Extensible Language Markup*), avec trois niveaux d'annotation : FTB-XML Texte Brut ; FTB-XML Constituants ; FTB-XML Fonctions. Le schéma XML est fondé sur la TEI (*Text Encoding*

Initiative), permettant ainsi de structurer les données par article, chaque balise <TEXT> correspondant à un extrait différent. Chaque extrait contient des métadonnées, structurées au moyen d'attributs de la TEI (comme pour la date ou le nom de l'auteur) et d'attributs spécifiques comme le domaine (« argument ») et l'identifiant unique de la phrase (« nb », voir figure 1).

Les déclinaisons « Constituants » et « Fonctions » de ce format d'origine sont également structurées en XML et sont les plus richement annotées. Pour les tokens, les balises <w> encadrent les annotations (catégorie, lemme, sous-catégorie, flexion), et sont enchâssées pour les composants de composés. Les constituants sont encadrés par une balise spécifiant leur nature (par exemple <PP> pour syntagme prépositionnel ou <NP> pour syntagme nominal). La fonction éventuelle du constituant est portée par un argument « fct » . Ainsi, dans la figure 1, le premier NP a la fonction Sujet (« SUJ »).

```
<text>
   <SENT argument="ETR" author="Minangoy Robert" date="1990-01-19"
      nb="1000" textID="456">
      <NP fct="SUJ">
         <w cat="PRO" ee="PRO-card-mp" ei="PROmp" lemma="six" mph="mp"
            subcat="card">Six</w>
         <PP>
            <w cat="P" compound="yes" ee="P" ei="P" lemma="d'entre">
               <w catint="P">d'</w>
               <w catint="P">entre</w>
            </w>
         <NP>
            <w cat="PRO" ee="PRO-3mp" ei="PRO3mp" lemma="eux" mph="3mp"
               subcat="pers">eux</w>
         </NP>
      </PP>
   </NP>
[...]
</text>
```

Figure 1. *Le format d'origine du FTB, avec métadonnées et fonctions*

3.1.2. *Le format TIGER-XML*

Afin de pouvoir l'interroger à l'aide de TIGERSearch, le corpus FTB a été structuré au format TIGER-XML (König et Lezius, 2000). Sans entrer dans les détails du format (Liégeois et Abeillé, 2018), les annotations des nœuds terminaux (les tokens) sont distinguées de celles des nœuds non terminaux, qui incluent les mots composés, les constituants et leurs fonctions.
L'outil TIGERSearch permet d'interroger l'ensemble des annotations, en les combinant éventuellement : lemmes et catégories des tokens, catégories et fonctions des constituants... Les résultats obtenus peuvent être visualisés sous la forme d'arbres syntaxiques (figure 2). En outre, TIGERSearch étant intégré à la version portail

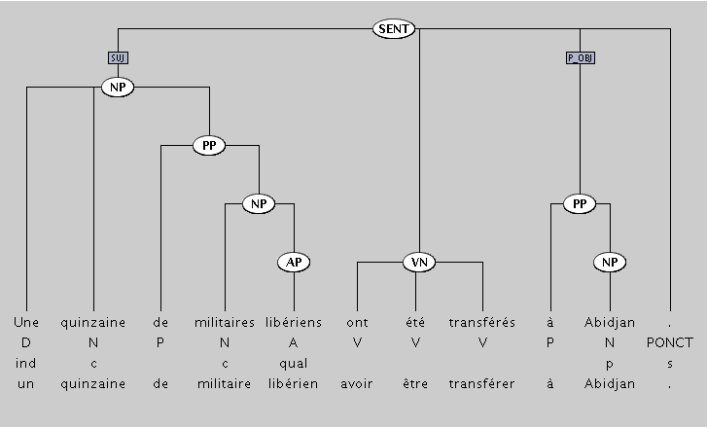

Figure 2. *Exemple d'affichage sous TIGERSearch*

de TXM (Heiden S., 2010), l'interrogation de l'ensemble des annotations du FTB est possible grâce à une plateforme en ligne sur le site du projet (`ftb.linguist.univ-paris-diderot.fr/`, section « Interroger »).

3.1.3. *Le format Penn TreeBank*

Le format Penn TreeBank (PTB) (Taylor *et al.*, 2003) caractérise la hiérarchie des constituants avec un système de parenthèses. Dans une phrase, une paire de parenthèses correspond à un niveau de l'arbre (exemple 5). Cette version du FTB utilise un jeu d'étiquettes simplifié pour les mots et des syntagmes, et certaines annotations sont perdues, comme le lemme ou la flexion, même si tous les composés sont gardés. Elle a toutefois l'avantage de permettre une visualisation graphique et l'interrogation du corpus à l'aide d'outils en ligne comme Tregex (Levy et Andrew, 2006) (figure 3).

(5) (SENT (NP-SUJ (D Une) (N quinzaine) (PP (P de) (NP (N militaires) (AP (A libériens))))) (VN (V ont) (V été) (V transférés)) (PP-P_OBJ (P à) (NP (N Abidjan))) (PONCT .))

3.1.4. *La version en dépendances (format CoNLL)*

Le quatrième format distribué est CoNLL, après conversion du corpus en arbres de dépendance (Candito *et al.*, 2010), selon le programme de conversion de Candito *et al.* (2010) légèrement modifié suite aux travaux de Seddah *et al.* (2013) pour la *SPMRL Shared Task*. La conversion est fondée sur le principe de définition d'une tête lexicale au sein de toute forme de syntagme. Les choix de conversion suivent les choix d'annotation syntaxique de la version en constituants (voir section 2.2), sauf que les

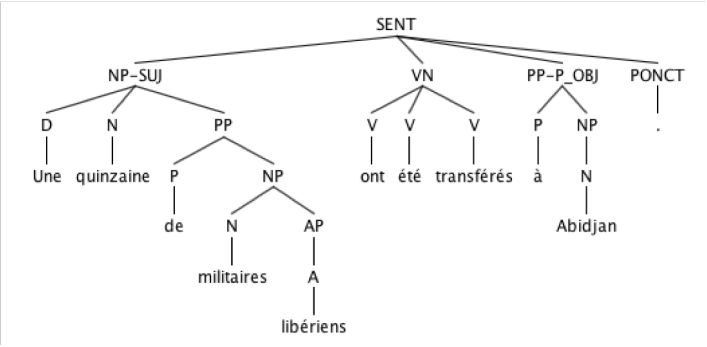

Figure 3. *Exemple d'affichage sous Tregex*

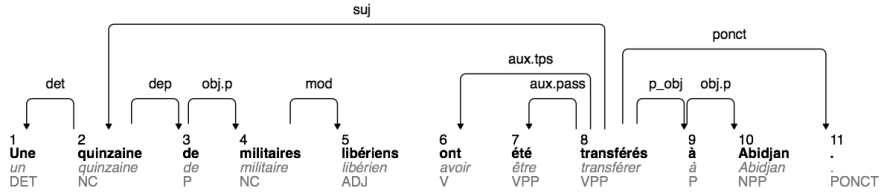

Figure 4. *Exemple d'affichage en dépendances*

prépositions sont toujours des têtes. Tous les mots composés ont été gardés, avec une étiquette de relation spéciale. Les fonctions étant associées aux mots et non aux syntagmes, seules ont été gardées les fonctions simples, et 12 fonctions ont été ajoutées automatiquement, en particulier pour les mots grammaticaux : *det* pour les déterminants, *obj-p* pour les compléments de préposition, *mod* pour les adjectifs épithètes et les relatives au sein du syntagme nominal, ou les adverbes au sein du syntagme adjectival, *dep* pour les autres dépendants des adjectifs et des noms, par exemple. Les auxiliaires de temps, passif et causatif reçoivent des étiquettes distinctes.
Cette version en dépendances (figure 4) est distribuée au format CoNLL, avec un token par ligne, et a donné lieu à un corpus dérivé au format "Universal dependencies" (Nivre *et al.*, 2016) (voir section 5.2).

3.2. *Le FTB (v.1.0) en quelques chiffres*

Nous présentons quelques chiffres concernant la distribution des mots, des catégories morphosyntaxiques, des syntagmes et des fonctions dans la version 1.0 du FTB.

3.2.1. *La répartition des catégories morphosyntaxiques*

Le corpus FTB comprend 644 595 tokens, signes de ponctuation inclus. Parmi ceux-ci, 553 024 sont des mots simples (avec trait « lemma » et/ou du trait « word »), pour un total de 30 688 types différents, et 78 634 des composants de composés. Le tableau 3 présente la distribution des catégories morphosyntaxiques pour les mots simples : certaines catégories sont surreprésentées (Nom, Préposition, Déterminant, Verbe) et d'autres, sous-représentées (Préfixe, Mot Étranger, Interjection). Si l'on compte en tokens, les mots les plus fréquents sont les Noms et les Prépositions, et les Clitiques sont d'un tiers plus nombreux que les autres Pronoms. Si l'on compte en types, les catégories les plus fréquentes sont les Noms et les Verbes, et les Pronoms sont 3,5 fois plus nombreux que les Clitiques.

Catégorie	Tokens	Types	Les 5 mots les plus fréquents
Nom	134 137	15 069	*pourcents, francs, M., milliards, millions*
Préposition	85 534	124	*de, à, des, d', du*
Déterminant	80 986	664	*la, l', les, le, un*
Ponctuation	79 862	18	, . ") (
Verbe	68 452	10 049	*est, a, ont, sont, été*
Adjectif	36 252	5 627	*français, deux, autres, dernier, premier*
Adverbe	22 510	731	*pas, plus, ne, n', aussi*
Conjonction	18 292	58	*et, que, ou, qu', mais*
Clitique	15 751	73	*s', se, il, on, en*
Pronom	10 461	234	*qui, dont, que, où, qu'*
Préfixe	423	58	*ex-, non-, vice-, quasi-, micro-*
Mot Étranger	299	208	*eiido, eiidos, bank, and, of*
Interjection	65	29	*hélas, bref, non, oui, attention*

Tableau 3. *Les catégories morphosyntaxiques dans le FTB (mots simples)*

Le tableau 4 présente la distribution des mots composés, selon leur catégorie morphosyntaxique : les plus fréquents sont les noms composés, et les adverbes composés, suivis de près par les prépositions complexes ou locutions prépositionnelles.

Pour savoir si ces chiffres sont représentatifs, nous manquons de données comparables pour d'autres corpus écrits, en particulier concernant les mots composés. L'on peut penser que la catégorie Clitique est plus représentée dans les corpus oraux, en raison de la fréquence des clitiques sujets (Blanche-Benveniste *et al.*, 1984), et la catégorie Interjection en registre informel.

3.2.2. *La répartition des syntagmes et des fonctions*

Si l'on considère la répartition des syntagmes dans le corpus (tableau 5), le syntagme nominal (NP) est surreprésenté, et les relatives (Srel) sont plus nombreuses que les autres subordonnées. Les moins fréquents sont les syntagmes adverbiaux (AdP),

Catégorie	Tokens	Exemples
Nom	14 398	*Côte-d'Ivoire, Médecins Sans Frontière*
Adverbe	5 819	*un peu, à grands frais*
Préposition	5 215	*avant de, faute de*
Déterminant	3 633	*la plupart de, trois cents*
Conjonction	1 322	*alors que, depuis que*
Verbe	878	*avoir affaire, prendre la fuite*
Pronom	574	*celle-ci, lui-même*
Adjectif	564	*hors-cadre, est-allemand*
Clitique	52	*l'on*
Etranger	34	*New-Deal, open-market*
Interjection	6	*au secours, oh là là*

Tableau 4. *Les catégories morphosyntaxiques dans le FTB (mots composés)*

ce qui est en partie lié au choix de ne pas annoter de syntagme quand l'adverbe est seul.

Constituant	Nombre	Exemples
Syntagme nominal	151 840	*le rapport Delors, cette intention*
Syntagme prépositionnel	82 620	*du même genre, de Bruxelles*
Noyau verbal	50 780	*on devrait, il est*
Syntagme adjectival	24 040	*précieux, social-démocrate*
Syntagme coordonné	15 659	*et de la retenue, ou non*
Syntagme infinitif	12 568	*de prouver cette intention*
Syntagme participial	8 284	*approuvées par le conseil européen*
Subordonnée relative	6 636	*qui est envisagée dans le rapport Delors*
Autre subordonnée	6 100	*combien celle-ci restait incertaine*
Incise et parenthèse	3 527	*leur état d'esprit a changé*
Syntagme adverbial	1 349	*plus tôt, un peu vite*

Tableau 5. *Les catégories des constituants dans le corpus*

Si l'on considère la répartition des fonctions (tableau 6), la fonction Sujet est la plus représentée, et la fonction Modifieur (MOD) est quasiment au même niveau. Les attributs du sujet (ATS) sont beaucoup plus fréquents que les attributs de l'objet (ATO), qui est la fonction la moins représentée du corpus. Les 19 combinaisons de fonctions (SUJ/OBJ, etc.) sont associées aux VN avec clitiques, et notent leur ordre aussi bien que leur fonction.

Fonction	Nombre	Exemples
SUJ	34 878	*le deutschemark, plusieurs voies*
MOD	34 747	*ici, lui-même*
OBJ	27 590	*cette intention, la construction monétaire*
ATS	6 228	*une voie praticable, négatives*
A-OBJ	4 570	*à l'économie allemande, à reculer*
DE-OBJ	4 013	*du gouvernement, de toutes parts*
P-OBJ	3 173	*par une partie des militants*
ATO	486	*comme des privilèges, reprendre ses droits*
SUJ/OBJ	284	*il y en a, je l'ai appelé*
SUJ/A-OBJ	139	*il m'apparaît, il leur apprend*
SUJ/DE-OBJ	53	*on s'en doute, nous en séparer*
SUJ/MOD	28	*j'y vois, on en connaît*
OBJ/A-OBJ	15	*l'y autorise, ne l'y obligerait*
SUJ/P-OBJ	14	*on y reste, j'y suis*
SUJ/ATS	11	*il l'a été, on l'était*
OBJ/SUJ	11	*en faut-il, les soumettront-ils*
OBJ/DE-OBJ	7	*l'en empêcha, l'en pressent*
A-OBJ/SUJ	5	*nous semble-t-il, se demande-t-il*
DE-OBJ/SUJ	5	*en pâtira-t-elle, en sera-t-il*
A-OBJ/DE-OBJ	4	*lui en est donnée, nous en rebattent*
A-OBJ/OBJ	4	*se le procurer, vous le diront*
SUJ/OBJ/A-OBJ	3	*nous le leur apprendrons*
A-OBJ/MOD	2	*leur y ont enseigné*
OBJ/MOD	2	*nous en a notifié, s'en faire*
OBJ/P-OBJ	1	*s'y présente*
SUJ/A-OBJ/OBJ	1	*on nous l'a toujours dit*
SUJ/MOD/OBJ	1	*on nous l'a changé*

Tableau 6. *Les fonctions grammaticales dans le corpus*

4. Une évaluation du corpus FTB

Une évaluation quantitative n'a pas été effectuée lors des différentes campagnes d'annotation. À chaque fois, les annotateurs travaillaient sur des sorties annotées automatiquement (voir section 2.1) en consultant des guides d'annotation dédiés, avec des réunions régulières pour trancher les cas difficiles ou inattendus (qui venaient compléter les guides). Qu'il s'agisse des mots composés, des étiquettes morphosyntaxiques, des constituants ou des fonctions, chaque fichier (de 500 phrases) était corrigé par un premier annotateur, puis par un second (généralement plus expérimenté) qui vérifiait et améliorait le résultat. Les annotateurs étaient tous des étudiants avancés en linguistique (L3 ou master à Paris 7 ou Paris 10) et certains sont restés plusieurs années dans le projet. Il n'a pas été question à l'époque de faire annoter en parallèle le même fi-

chier et de mesurer l'accord interannotateur, comme on le ferait aujourd'hui (Artstein et Poesio, 2008). Pour combler cette lacune, nous avons réalisé une évaluation *a posteriori* en prenant la version finale distribuée comme référence.

4.1. *Une évaluation morphosyntaxique*

Une qualité du FTB est la richesse de ses annotations morphosyntaxiques, qui incluent de nombreuses sous-catégories ainsi que toutes les informations flexionnelles : ainsi *les* est annoté féminin ou masculin selon le nom qui suit, si c'est un Déterminant, et selon son antécédent si c'est un Clitique. Nous avons pris 100 phrases au hasard. En excluant les ponctuations (qui ne posent pas de problème d'annotation) et les composants de composés (qui ont des catégories plus sommaires), elles comprennent 2 168 tokens. Nous n'avons plus les fichiers de sortie de l'époque, donc nous avons pris une version sans annotation préalable, sauf les mots composés, ce qui est une tâche plus difficile qu'à l'époque. Un expert linguiste les a annotés (un mot par ligne), en utilisant les guides et le jeu des 122 étiquettes internes (attribut ei). En comparant avec le FTB, la valeur du kappa (Cohen, 1960) est 0,97 (pour un pourcentage d'accord de 97,3 %). Les désaccords concernent surtout le genre des Clitiques, des Pronoms et des Noms propres, difficile à renseigner hors contexte sans marque morphologique (*je*, *Air Inter*). Ils concernent aussi Adjectif ou Participe passé, Adverbe ou Préposition, Déterminant, Adjectif ou Pronom pour certains indéfinis (*l'un*, *l'autre*), *de* comme Préposition ou Déterminant. L'adjudication donne raison au corpus, selon les guides d'annotation. Si l'on se limite aux catégories lexicales (attribut pos), l'accord est quasi parfait (κ = 0,99, pour un pourcentage d'accord de 99 %), sur 9 catégories car Interjection, Mot étranger et Préfixe n'étaient pas représentés dans l'échantillon.

4.2. *Une évaluation en constituants*

Afin d'évaluer les constituants, nous avons pris les mêmes 100 phrases et demandé à l'expert de les annoter (format PTB) selon les guides du corpus. Il s'agit d'une tâche sans annotation automatique préalable, donc plus difficile que pour les annotateurs du projet. L'expert a annoté les principaux constituants, c'est-à-dire tous les noyaux verbaux et syntagmes verbaux, toutes les propositions (relatives, subordonnées, internes) et tous les syntagmes coordonnés, mais seulement les syntagmes nominaux, adjectivaux, adverbiaux et prépositionnels majeurs, c'est-à-dire recevant une fonction. Le fichier obtenu comporte 925 syntagmes, et nous l'avons comparé au FTB, avec le programme *evalb* (Sekine *et al.*, 2008), et obtenu une FMesure de 89,65. Les principaux cas de désaccord portaient sur le nombre de syntagmes (18 en plus ou en moins), soit un taux d'accord de 98,1%, sur les bornes ouvrantes (taux d'accord de 99,8 %), sur les bornes fermantes (taux d'accord de 97,5 %), et sur les étiquettes (taux d'accord de 99,1 %). Les désaccords sur les étiquettes portaient sur AP et VPpart ou NP et PP et étaient liés aux désaccords sur les catégories lexicales. Les désaccords sur les bornes

fermantes concernent l'annotation des appositions (dans le NP ou en dehors), et des coordinations.

4.3. *Une évaluation des fonctions*

Pour les fonctions grammaticales, nous avons reproduit la situation des annotateurs de l'époque, car nous avons gardé certaines sorties de l'annotateur fonctionnel automatique. Nous avons pris 100 phrases au hasard. Un expert linguiste a corrigé les fonctions associées aux 616 syntagmes concernés. Par rapport au FTB, κ est à 0,91 (pour un pourcentage d'accord de 93,2 %). Les principaux désaccords concernent les syntagmes prépositionnels, ajouts (MOD) ou compléments (A-OBJ, DE-OBJ, P-OBJ) et les fonctions associées aux Clitiques (pas de fonction pour les Clitiques figés). L'expert avait oublié la fonction MOD pour les constituants disloqués. Comme il ne pouvait pas corriger les syntagmes, contrairement aux annotateurs de l'époque, il y a quelques discordances (31 fonctions en plus ou en moins entre les deux versions), la version distribuée ayant fait l'objet de post-corrections en syntagmes supplémentaires (κ est à 0,86 si l'on en tient compte, pour un pourcentage d'accord de 88,9 %).

Annotation	Tokens	Nombre d'étiquettes	Taux d'accord (kappa)	Pourcentage d'accord
cat. lexicales	2 168	11	0,99	99 %
souscat + morpho	2 168	122	0,97	97,3 %
syntagmes	925	10	0,99	98,1 %
fonctions	616	11	0,91	93,2 %

Tableau 7. *Les taux d'accord sur les étiquettes du FTB (sur 100 phrases)*

Nous concluons que les annotations du FTB sont de très bonne qualité, même s'il reste toujours des erreurs qui sont corrigées régulièrement, et que les choix d'annotation sont reproductibles.

5. Exemples d'utilisation du FTB

Depuis sa création, le corpus FTB a été utilisé pour des applications variées, dans le cadre d'études linguistiques ou psycholinguistiques, ou en traitement automatique des langues.

5.1. *Utilisation du FTB pour des études linguistiques*

Le corpus FTB a été exploité pour réaliser diverses études linguistiques. Sans être exhaustif, nous pouvons citer les études sur les phrases sans verbe de Laurens (2008) et les relatives sans verbe de Bîlbîie et Laurens (2009), les coordinations itératives

(Mouret, 2005), (Mouret, 2007), les coordinations de phrases avec ellipse (Abeillé et Mouret, 2010), les relatives en *dont* (Abeillé *et al.*, 2016) et l'inversion du sujet dans les relatives en *que* (Pozniak *et al.*, 2019). Ces études s'appuient sur les annotations réalisées au niveau des catégories morphosyntaxiques et des types de constituants, mais aussi sur l'annotation fonctionnelle. Abeillé *et al.* (2016) ont trouvé que dans la majorité des cas, *dont* est utilisé pour le complément du sujet (dont la majorité, dont le directeur). En comparant avec un grand corpus littéraire (Frantext aux XIXe et XXe siècles), Abeillé et Winckel (2018) ont trouvé une proportion du même ordre, ce qui est un indice de la représentativité du FTB pour les questions de syntaxe, du moins à l'écrit en registre formel.

À l'aide de TIGERSearch, nous avons cherché les catégories associées à la fonction Sujet, la plus fréquente dans le corpus. Sans surprise, les syntagmes nominaux sont les plus nombreux avec 26 789 occurrences au total, soit près de 77 % (voir tableau 8), suivis par les Clitiques (dans ce cas, la fonction Sujet est portée par le noyau verbal (VN)). Quand un syntagme coordonné (COORD) porte la fonction, il s'agit d'une coordination itérative (*ni la France ni l'Angleterre*). Au total, 5 types de constituants sont employés comme sujet. Des requêtes plus fines permettent par exemple d'étudier l'inversion du sujet ou l'accord sujet verbe. Ainsi, Pozniak *et al.* (2019) ont trouvé que 50 % des sujets nominaux sont inversés dans les relatives en *que*. Nous avons trouvé 6 cas d'infinitifs sujets inversés (sur 99), dont 5 avec *vaut / vaudrait mieux* :

(6) a. `[...]` *mieux vaut* [*s'entraîner* $_{VPinf\text{-}SUJ}$].
b. *Reste* [*à savoir quelle en sera l'ampleur* $_{VPinf\text{-}SUJ}$].

Mouret (2007) a trouvé que les sujets infinitifs coordonnés permettent aussi bien l'accord singulier que l'accord pluriel, et ce dans deux phrases d'un même article :

(7) a. [*Ne pas savoir se servir d'un ordinateur (66,5 %), travailler à temps réduit (64,9 %) et être âgé de plus de quarante-cinq ans (52,3 %)* $_{VPinf\text{-}SUJ}$] *constitue un handicap, lors d'une promotion.*
b. *À l'inverse,* [*être un homme (48,6 %) et avoir des diplômes (85,6 %)* $_{VPinf\text{-}SUJ}$] *sont manifestement des atouts* `[...]`.

Constituant	Sujet	Exemples
NP	26 789	*lequel, Mr Hans Modrow*
VN	7 946	*on devrait, il y avait*
VPinf	99	*assister à un tel bouleversement*
Ssub	24	*qu'il soit employé comme nom ou comme adjectif*
COORD	20	*ni système de prix ni marché*

Tableau 8. *La fonction Sujet dans le corpus*

5.2. *L'utilisation du FTB pour la constitution d'autres ressources*

Le corpus FTB a permis la création de ressources dérivées.

5.2.1. *Les lexiques dérivés*

Un certain nombre de lexiques ont été dérivés comme TreeLex (Kupść, 2009 ; Kupść et Abeillé, 2008). Ce dictionnaire de valence, dont les entrées on été extraites automatiquement du corpus, comporte les adjectifs (2 200 entrées) et les verbes (2 000 entrées) du corpus. La valence, automatiquement extraite du corpus, convertie pour le passif, le réfléchi, etc. et corrigée manuellement, est indiquée pour chacun des lemmes. Un dictionnaire de valence d'adjectifs a été étendu par Fabre et Kupść (2009). Le lexique Nomage (Balvet *et al.*, 2011) contient quant à lui un ensemble de noms déverbaux (morphologiquement dérivés d'un verbe) pour lesquels, à partir des exemples présents dans le corpus FTB, les auteurs ont ajouté une couche d'annotation sémantique (arguments et classe aspectuelle).

5.2.2. *Les corpus dérivés*

Plusieurs corpus ont été dérivés à partir des versions antérieures du FTB, par exemple le corpus de Dublin (2007) qui comporte 4 741 phrases du FTB (*Modified FTB*) (Schluter et van Genabith, 2007), le corpus d'Aix-en-Provence qui comporte 1 471 phrases du FTB, en ajoutant des annotations des grammaires de propriétés (FTB-LPL) (Blache et Rauzy, 2012). Le Dependency Corpus d'Alpage (FTB-DEP) comporte 12 500 phrases du FTB converties au format CoNLL (Candito *et al.*, 2009). Un corpus de référence (*gold*) a été produit par Seddah *et al.* (2013) pour la *Shared Task* de SPRML 2013 : 38 fichiers du FTB ont été convertis au format CoNLL. Enfin, le corpus U-FTB a été obtenu après après conversion automatique (Seddah *et al.*, 2018) en suivant les étiquettes "Universal Dependencies" (http ://universaldependencies.org/) (McDonald *et al.*, 2013).

D'autres projets ont ajouté des annotations supplémentaires : ainsi Sagot *et al.* (2012) y ajoutent les entités nommées, Candito et Seddah (2012a), Ribeyre *et al.* (2014) ajoutent des relations de syntaxe profonde, pour les dépendances à distance, le passif, les constructions impersonnelles, etc., Djemaa *et al.* (2016) ajoutent des informations sémantiques (de type framenet) sur un sous-ensemble de prédicats, et Danlos *et al.* (2015) ajoutent un étiquetage des connecteurs de discours et des relations de discours (French Discourse Treebank) sur l'ensemble du corpus.
Du côté de la psycholinguistique, Pynte *et al.* (2009) ont ajouté les temps de lecture et les mouvements oculaires sur 52 173 tokens du FTB (qui constituent la partie française du Dundee Corpus), et (Rauzy et Blache, 2012) ont annoté avec temps de lecture et mouvements oculaires 198 phrases (6 572 tokens) du FTB (corpus physiologique du LPL). Hale (2014) a, quant à lui, calculé un modèle de surprise fondé sur le FTB, très utilisé en psycholinguistique computationnelle.

Les guides d'annotation ont par ailleurs été réutilisés dans le projet d'évaluation et d'annotation Easy (Paroubek *et al.*, 2007), pour les corpus Passage (Villemonte de La Clergerie *et al.*, 2008) et Sequoia (Candito *et al.*, 2014), ainsi que plus récemment pour des corpus de questions (Seddah et Candito, 2016), de médias sociaux (Seddah *et al.*, 2012), ou un corpus oral de radio (Abeillé et Crabbé, 2013), avec des adaptations nécessaires.

6. Comparaison avec d'autres corpus français annotés

6.1. *Comparaison avec des corpus taggés*

Depuis la création du FTB, de nombreux corpus taggés ont vu le jour. Les plus gros, pour le français écrit, sont le frWaC (2,6 milliards de mots) et le frTenTen (10 milliards de mots) issus de pages Web (Baroni *et al.*, 2009). Ils utilisent un jeu d'étiquettes réduit (33 étiquettes). Ils sont annotés automatiquement avec Treetagger, sans correction, ce qui engendre de nombreuses erreurs. Il en va de même du corpus littéraire Frantext (253 millions de mots), entièrement catégorisé depuis 2018, qui utilise un jeu de 22 étiquettes, récemment mis à jour pour correspondre au jeu de la version en dépendances du FTB (Candito *et al.*, 2009).

6.2. *Comparaison avec d'autres corpus arborés*

La plupart des corpus arborés qui ont vu le jour depuis, se sont inspirés du FTB, et ont eu pour but d'être plus équilibrés et d'ajouter des relations pour prendre en compte l'oral ou pour être plus proches de la sémantique. Les corpus écrits Passage et Séquoia sont plus équilibrés que le FTB puisque, outre des textes journalistiques (*Est Républicain*), ils incluent aussi des textes de Wikipédia, du Parlement européen et de l'Agence européenne des médicament. Le corpus Passage (Villemonte de La Clergerie *et al.*, 2008) comprend 2 millions de mots, dont 4 000 phrases corrigées à la main. Le corpus Séquoia (Candito et Seddah, 2012b) comporte 3 204 phrases et 69 246 tokens, annotés automatiquement et corrigés, disponibles aux formats PTB ou CoNLL, avec un jeu de 28 étiquettes lexicales. Pour les syntagmes, et pour les dépendances, il utilise le même jeu d'étiquettes que le FTB (format PTB et format CoNLL). Dans sa dernière version (Candito *et al.*, 2014), ont été ajoutées des annotations « profondes » plus proches de la sémantique, pour le passif ou l'impersonnel, le sujet implicite des infinitifs ou les cas d'ellipse. Le nombre de relations est donc plus important avec 28 étiquettes différentes.

Le corpus French GSD (Google Stanford Dependencies) a été créé en 2015 au sein du projet multilingue de Universal Dependency Treebanks (McDonald *et al.*, 2013). Il a été ensuite modifié pour se conformer au schéma d'annotation Universal Dependencies (Nivre *et al.*, 2016), et est intégré aux différentes versions du projet Universal Dependencies, depuis la version 2.0. Il comporte 16 342 phrases (389 363 tokens), à partir de blogs et de pages Wikipédia, sans métadonnées. Il s'appuie sur 17 étiquettes

lexicales universelles, enrichies par 37 traits morphosyntaxiques. Les clitiques ne sont pas distingués des pronoms forts. Les phrases sont segmentées de telle sorte que les seuls tokens contenant des espaces sont des nombres. Des expressions polylexicales ont été annotées (essentiellement des mots composés grammaticaux) en utilisant une représentation plate, et une étiquette de relation spécifique. Pour la syntaxe, il s'appuie sur un jeu de 34 relations universelles, avec 16 sous-types. Le choix est de toujours avoir pour tête une catégorie majeure (ainsi les prépositions ne sont jamais têtes, ni le verbe *être*). Pour les groupes prépositionnels dépendant de verbes, il ne distingue pas systématiquement complément et modifieur, ni objet et attribut pour les infinitifs ou les complétives. Il a fait l'objet de corrections manuelles et automatiques, mais n'a pas été évalué pour les catégories lexicales ni les traits morphosyntaxiques. Pour une évaluation des relations sur 100 phrases, voir (Guillaume *et al.*, 2019).
Le Corpus d'Étude pour le Français Contemporain (CEFC) (Debaisieux *et al.*, 2016), qui compte 10 millions de mots dont 6 millions de textes écrits, se veut représentatif des variétés du français oral et écrit. Il mêle des transcriptions et des textes variés (interviews, oral en interaction, oral spontané, parole avec variation régionale, parole d'enfant, documents administratifs, journaux, romans). Seule une sous-partie (172 000 mots) est annotée pour la syntaxe, selon un schéma en dépendances, et corrigée manuellement. Elle utilise 21 étiquettes lexicales, et 12 relations fonctionnelles, 9 pour la « microsyntaxe » (par exemple sujet, spécifieur et dépendant, qui ne distinguent pas entre complément et modifieur) et 3 pour la « macrosyntaxe » (par exemple périphérique et parenthétique).

Corpus	Tokens	Tokens corrigés	Étiquettes mots	Étiquettes syntagmes	Relations
FTB	644k	644k	218	12	8/20
Sequoia	69k	69k	28	12	8/28
CEFC	10M	172k	21	-	12
U-GSD	389k	389k	17/74	-	50

Tableau 9. *Les principaux corpus arborés pour le français*

Le FTB reste une ressource unique par la finesse de ses annotations lexicales, comme par le volume de ses données corrigées.

7. Conclusion

Le Corpus arboré du français (French Treebank) de l'université Paris Diderot a joué un rôle majeur dans l'outillage du français, et a inspiré de nombreux projets dérivés. C'est un corpus homogène, par le choix des textes (journal *Le Monde*) et par leurs dates (1990-1993). Il est de taille moyenne comparé à certains corpus annotés plus récents, et non équilibré, mais avec une richesse d'annotation inégalée, tant par le jeu d'étiquettes (218 étiquettes morphosyntaxiques, 20 étiquettes syntaxiques), que par le nombre de mots composés. Les nombreuses phases de validation manuelle en

font une ressource de qualité, même si des erreurs résiduelles peuvent toujours subsister, et que certains choix d'annotation sont aujourd'hui datés. Sa version finale, qui est disponible en plusieurs formats, permet à la fois des utilisations variées en traitement automatique des langues (format en dépendances CoNLL par exemple) et des requêtes à l'aide d'outils génériques (TXM, TIGERSearch, Tregex) pour des utilisateurs linguistes. Disponible sur un site dédié, il est toujours d'actualité comme en témoigne le nombre croissant de nouveaux utilisateurs. Il n'a pas la prétention d'être représentatif des variétés du français, mais présente l'avantage de permettre des études syntaxiques et lexicales fondées sur « le bon usage » du français écrit contemporain.

Remerciements

Le projet a été soutenu par l'IUF, chaire junior 1996-2001 et chaire senior 2007-2012, d'Anne Abeillé, par l'université Paris Diderot, le LLF, le CNRTL et la DGLFLF. Nous tenons à remercier les relecteurs de TAL ainsi que, pour leur travail sur les premières versions du corpus, Nicolas Barrier (annotation en fonctions), Martine Cheradame (coordination des stagiaires et guides d'annotation), Alexandra Kinyon, Jacques Steinlin et François Toussenel (annotation en constituants), Rodrigo Reyes (annotation morphosyntaxique) et pour leur contribution à la version finale : Marie-Hélène Candito et Benoit Crabbé (détection d'erreurs, conversion au format CoNLL, annotation en fonctions), Vanessa Combet (correction), Achille Falaise, Clément Plancq, Alexandre Roulois et Johan Ferguth (site du projet, plateforme d'interrogation et maintenance des différents formats).

8. Bibliographie

Abeillé A., *Corpus arboré pour le français : guide d'annotation en fonctions*, Université Paris Diderot, Paris, 2004.

Abeillé A., « Les syntagmes conjoints et leurs fonctions syntaxiques », *Langages*, vol. 160, p. 42-66, 2005.

Abeillé A., Barrier N., « Enriching a French treebank », *4th LREC*, Lisbonne, p. 2233-2236, 2004.

Abeillé A., Clément L., *Corpus arboré pour le français : guide d'annotation morphosyntaxique*, Université Paris Diderot, Paris, 1999a.

Abeillé A., Clément L., « A reference tagged corpus for French », *in* T. Brants H. U. (dir.), *LINC, EACL*, Bergen, p. 17-24, 1999b.

Abeillé A., Clément L., Toussenel F., « Building a Treebank for French », *in* Abeillé A. (dir.), *Treebanks : Building and Using Parsed Corpora*, Kluwer, Dordrecht, p. 165-188, 2003.

Abeillé A., Crabbé B., « Vers un treebank du français parlé », *20ème Conférence TALN*, 2013.

Abeillé A., Godard D., « La position de l'adjectif épithète en français : le poids des mots », *Recherches linguistiques de Vincennes*, vol. 28, p. 9-32, 1999.

Abeillé A., Godard D., « A Class of 'lite' Adverbs in French », *in* Camps J., Wiltshire C. (dir.), *Romance Syntax, Semantics and their L2 Acquisition*, John Benjamins, p. 9-25, 2001.

Abeillé A., Hemforth B., Winckel E., « Les relatives en dont : études empiriques », *5ème CMLF*, ILF, Tours, p. 26-43, 2016.

Abeillé A., Kinyon A., Clément L., « Building a treebank for French », *2d LREC*, Athènes, 2000.

Abeillé A., Mouret F., « Quelques contraintes sémantiques et discursives sur les coordinations elliptiques », *Revue de sémantique et de pragmatique*, vol. 24, nº 3, p. 177-206, 2010.

Abeillé A., Toussenel F., Chéradame M., *Corpus arboré pour le français : guide d'annotation en constituants*, Université Paris Diderot, Paris, 1999.

Abeillé A., Winckel E., « Dont and de qui relatives in written French », *Grammar and Corpora*, 2018.

Artstein R., Poesio M., « Inter-Coder Agreement for Computational Linguistics », *Computational Linguistics*, vol. 34, nº 4, p. 555-596, 2008.

Balvet A., Barque L., Condette M.-H., Haas P., Huyghe R., Marin R., Merlo A., « La ressource Nomage. Confronter les attentes théoriques aux observations du comportement linguistique des nominalisations en corpus », *TAL*, vol. 52, nº 3, p. 129-152, 2011.

Baroni M., Bernardini S., Ferraresi A., Zanchetta E., « The WaCky wide web : a collection of very large linguistically processed web-crawled corpora », *Language ressources and evaluation*, vol. 43, nº 3, p. 209-226, 2009.

Blache P., Rauzy S., « Enrichissement du FTB : un treebank hybride constituants/propriétés », *19ème Conférence TALN*, 2012.

Blanche-Benveniste C., Deulofeu J., Stéfanini J., van den Eynde K., *Pronom et syntaxe : l'approche pronominale et son application au français*, SELAF, 1984.

Brill E., A corpus-based approach to language learning, Thèse de Doctorat, University of Pennsylvania, 1993.

Bîlbîie G., Laurens F., « A Construction-based Analysis of Verbless Relative Adjuncts in French and Romanian », *in* Müller S. (dir.), *Proceedings 16th HPSG Conference*, CSLI Publications, Stanford, p. 5-25, 2009.

Candito M.-H., Crabbé B., Denis P., « Statistical French dependency parsing : treebank conversion and first results », *7th LREC*, La Valletta, 2010.

Candito M.-H., Crabbé B., Denis P., Guérin F., « Analyse syntaxique du français : des constituants aux dépendances », *16ème Conférence TALN*, Senlis, 2009.

Candito M.-H., Perrier G., Guillaume B., Ribeyre C., Fort K., Seddah D., de la Clergerie E., « Deep Syntax Annotation of the Sequoia French Treebank », *9th LREC*, Reykjavik, 2014.

Candito M.-H., Seddah D., « Effectively long-distance dependencies in French : annotation and parsing evaluation », *TLT11*, 2012a.

Candito M.-H., Seddah D., « Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical », *19e Conférence TALN*, Grenoble, France, 2012b.

Clément L., Construction et exploitation d'un corpus syntaxiquement annoté pour le français, Thèse de Doctorat, Université Paris 7, 2001.

Cohen J., « A Coefficient of Agreement for Nominal Scales », *Educational and Psychological Measurement*, vol. 20, nº 1, p. 37-46, 1960.

Danlos L., Colinet M., Steinlin J., « FDTB1 : Repérage des connecteurs de discours dans un corpus français », *Discours*, 2015.

Debaisieux J.-M., Benzitoun C., Deulofeu H.-J., « Le projet ORFEO : Un corpus d'études pour le français contemporain », *Revue Corpus*, vol. 15, p. 91-114, 2016.

Djemaa M., Candito M.-H., Muller P., Vieu L., « Corpus annotation within the French FrameNet : a domain-by-domain methodology », *LREC*, 2016.

Fabre C., Kupść A., « Large and noisy vs small and reliable : combining 2 types of corpora for adjective valence extraction », *5th Corpus Linguistics conference*, Liverpool, 2009.

Fradin B., *Nouvelles approches en morphologie*, PUF, 2003.

Gross M., « Sur quelques groupes nominaux complexes », *in* Chevalier J.-C., Gross M. (dir.), *Méthodes en grammaire française*, Klincksieck, Paris, p. 97-119, 1976.

Guillaume B., de Marneffe M.-C., Perrier G., « Conversion et amelioration de corpus du francais annotes en Universal Dependencies », *TAL*, vol. 60, nº 2, p. 42-66, 2019.

Habert B., « Outiller la linguistique : de l'emprunt de techniques aux rencontres de savoirs », *Revue française de linguistique appliquée*, vol. IX, nº 1, p. 5-24, 2004.

Hale J. T., *Automaton Theories of Human Sentence Comprehension*, University of Chicago Press, 2014.

Heiden S. Magué J.-P. P. B., « TXM : Une plateforme logicielle open-source pour la textométrie conception et développement », *in* Bolasco S. (dir.), *Proc. 10th JADT*, p. 1021-1032, 2010.

Illouz G., Habert B., Folch H., Prévost S., « TyPex : Generic Feature for Text Profiler », *Computer-Assisted Information Retrieval*, RIAO, 2000.

Kinyon A., « A Language-Independent Shallow-Parser Compiler », *39th ACL and 10th EACL,Toulouse*, p. 322-329, 2001.

Kupść A., « TreeLex Meets Adjectival Tables », *International Conference RANLP*, Borovets (Bulgarie), 2009.

Kupść A., Abeillé A., « Treelex : a subcategorization lexicon automatically extracted from a French Treebank », *ICGL*, Workshop on syntactic annotations, Hong-Kong, 2008.

König E., Lezius W., The TIGER language A Description Language for Syntax Graphs, Formal Definition, Technical report, IMS, University of Stuttgart, 2000.

Laurens F., « French predicative verbless utterances », *in* Müller S. (dir.), *Proceedings 15th HPSG Conference, Keihanna*, CSLI Publications, Stanford, p. 152-172, 2008.

Levy R., Andrew G., « Tregex and Tsurgeon : tools for querying and manipulating tree data structures », *5th LREC*, 2006.

Liégeois L., Abeillé A., *Corpus arboré pour le français : guide d'interrogation*, Université Paris Diderot, Paris, 2018.

Maurel D., Belleil C., « Un dictionnaire électronique relationnel des noms propres liés à la géographie », *LINX*, vol. 34-35, p. 77-88, 1996.

McDonald R., Nivre J., Quirmbach-Brundage Y., Goldberg Y., Das D., Ganchev K., Hall K., Petrov S., Zhang H., Täckström O., Bedini C., Castell N. B., Lee J., « Universal dependency annotation for multilingual parsing », *51st ACL Meeting*, Sofia, p. 9297, 2013.

Miller P. H., *Clitics and Constituents in Phrase Structure Grammar*, Garland, 1992.

Mouret F., « La syntaxe des coordinations corrélatives du français », *Langages*, vol. 39, nº 160, p. 67-92, 2005.

Mouret F., Grammaire des constructions coordonnées. Coordinations simples et coordinations à redoublement en français contemporain, Thèse de Doctorat, Université Paris Diderot, 2007.

Nazarenko A., Habert B., Salem A., *Les linguistiques de corpus*, Armand Colin, 1997.

Nivre J., de Marneffe M.-C., Ginter F., Goldberg Y., Hajic J., Manning C. D., McDonald R. T., Petrov S., Pyysalo S., Silveira N., Tsarfaty R., Zeman D., « Universal Dependencies v1 : A Multilingual Treebank Collection », *10th LREC*, Portoroz, 2016.

Paroubek P., Vilnat A., Robba I., Ayache C., « Les résultats de la campagne EASY d'évaluation des analyseurs syntaxiques du français », *14ème Conférence TALN*, 2007.

Pozniak C., Abeillé A., Hemforth B., « French relatives and subject inversion : what's your preference ? », *in* Crysmann B., Sailer M. (dir.), *One-to-Many Relations in Morphology, Syntax and Semantics*, Language Science Press, Berlin, 2019.

Pynte J., New B., Kennedy A., « On-line syntactic and semantic influences in reading revisited », *Journal of Eye Movement Research*, vol. 3, n° 1, p. 1-12, 2009.

Rauzy S., Blache P., « Robustness and processing difficulty models. A pilot study for eye-tracking data on the French Treebank », *COLING*, 2012.

Ribeyre C., Candito M.-H., Seddah D., « Semi-Automatic Deep Syntactic Annotations of the French Treebank », *TLT13*, Tubingen, 2014.

Sagot B., Richard M., Stern R., « Annotation référentielle du Corpus Arboré de Paris 7 en entités nommées », *19ème Conférence TALN*, vol. 2, Grenoble, p. 535-542, 2012.

Schluter N., van Genabith J., « Preparing, Restructuring and Augmenting a French Treebank : Lexicalised Parsing or Coherent Treebanks », *10th PACLING*, Melbourne (Australie), 2007.

Seddah D., Candito M.-H., « Hard Time Parsing Questions : Building a QuestionBank for French », *10th LREC*, 2016.

Seddah D., Clergerie E. D. L., Sagot B., Alonso H. M., Candito M., « Cheating a Parser to Death : Data-driven Cross-Treebank Annotation Transfer », *11th LREC*, ELRA, Miyazaki, Japan, 2018.

Seddah D., Sagot B., Candito M.-H., Mouilleron V., Combet V., « The French social media bank : a treebank of noisy user generated content », *COLING*, Mumbai, 2012.

Seddah D., Tsarfaty R., Kübler S., Candito M.-H., Choi J. D., Farkas R., Foster J., Goenaga I., Gojenola K., Goldberg Y., Green S., Habash N., Kuhlmann M., Maier W., Nivre J., Przepiórkowski A., Roth R., Seeker W., Versley Y., Vincze V., Woli M., Wróblewska A., de la Clergerie E. V., « Overview of the SPMRL 2013 Shared Task : Cross-Framework Evaluation of Parsing Morphologically Rich Languages », *4th Workshop on Statistical Parsing of Morphologically Rich Languages*, ACL, Seattle, p. 146-182, 2013.

Sekine S., Collins M., Brooks D., Ellis D., *Evalb software*, 2008. `nlp.cs.nyu.edu/evalb`.

Silberztein M., *Dictionnaires électroniques et analyse automatique de textes : le système Intex*, Masson, Paris, 1993.

Taylor A., Marcus M., Santorini B., « Treebanks : Building and using parsed corpora », *in* Abeillé A. (dir.), *Treebanks*, Kluwer, Dordrecht, p. 5-22, 2003.

Toussenel F.*, « Marquage de constituants sur un corpus français, résultats et exploitation linguistiques »*, Mémoire de Master, Université Paris Diderot, 2001.

Villemonte de La Clergerie É., Hamon O., Mostefa D., Ayache C., Paroubek P., Vilnat A., « PASSAGE : from French Parser Evaluation to Large Sized Treebank », *6th LREC*, Marrakech, 2008.

Redonner du sens à l'accord interannotateur : vers une interprétation des mesures d'accord en termes de reproductibilité de l'annotation

Dany Bregeon[*,***] — **Jean-Yves Antoine**[*] — **Jeanne Villaneau**[**] — **Anaïs Halftermeyer**[***]

[*] *LIFAT (EA 6300), ICVL, Université de Tours*
jean-yves.antoine@univ-tours.fr
[**] *IRISA (UMR 6074) D6-Expression,Université de Bretagne Sud*
jeanne.villaneau@univ-ubs.fr
[***] *LIFO, ICVL, Université d'Orléans*
dany.bregeon@etu.univ-orleans.fr, anais.halftermeyer@univ-orleans.fr

RÉSUMÉ. *Les mesures d'accord interannotateur sont utilisées en routine par le TAL pour évaluer la fiabilité des annotations de référence. Pourtant, les seuils de confiance liés à cette estimation relèvent d'opinions subjectives et n'ont fait l'objet d'aucune expérience de validation dédiée. Dans cet article, nous présentons des résultats expérimentaux sur données réelles ou simulées qui visent à proposer une interprétation des mesures d'accord en termes de stabilité de la référence produite, sous la forme d'un taux moyen de variation de la référence entre différents groupes d'annotateurs.*

ABSTRACT. *Inter-coders agreement measures are used to assess the reliability of annotated corpora in NLP. Now, the interpretation of these agreement measures in terms of reliability level relies on pure subjective opinions that are not supported by any experimental validation. In this paper, we present several experiments on real or simulated data that aim at providing a clear interpretation of agreement measures in terms of the level of reproductibility of the reference annotation with any other set of coders.*

MOTS-CLÉS : *accord interannotateur, reproductibilité, niveau de fiabilité.*

KEYWORDS: *inter-coders agreement, reproductibility, reliability level.*

TAL. Volume 60 – n°2/2019, pages 45 à 69

1. Introduction

Utilisant de manière intensive des techniques d'apprentissage automatique entraînées sur des corpus, le traitement automatique des langues (TAL) a un besoin de plus en plus insatiable de ressources langagières massives. Face à ces besoins toujours croissants, le TAL a désormais recours fréquemment à ces corpus annotés automatiquement. La masse de données alors produite constitue une contrepartie intéressante du biais introduit, du moins si l'on fait le pari que les systèmes n'apprendront pas ce biais. Le recours à des ressources de qualité, annotées ou révisées manuellement, reste toutefois toujours pertinent, soit que l'on ne dispose pas de solution d'annotation automatique efficace, soit que la qualité des données d'apprentissage est un plus pour l'application visée.

La qualité des données annotées va au-delà des erreurs observées par rapport au guide d'annotation. Dans le cas de tâches complexes et/ou soumises à une forte subjectivité (pensons par exemple à la détection d'émotion), cette qualité répond au contraire avant tout à une exigence de fiabilité, c'est-à-dire de reproductibilité de l'annotation. Un corpus ne pourra en effet être considéré comme fiable et représentatif de la tâche considérée que si les annotations obtenues sont reproductibles par d'autres annotateurs que ceux choisis à l'initial (à l'idéal).

Dans le cas de corpus annotés par plusieurs personnes, on estime cette reproductibilité en observant l'accord qui existe entre chaque annotation individuelle. Le principe sous-jacent ici est que plus les annotateurs s'accordent entre eux, plus il y a de chances qu'ils se seraient également accordés avec n'importe quels autres annotateurs. Les chances que l'annotation soit reproductible sont donc d'autant plus élevées que l'accord entre les annotateurs l'est également.

Plusieurs métriques d'évaluation de l'accord interannotateur ont été proposées dans la littérature et sont utilisées en routine en TAL. Les plus répandues dans notre communauté sont ainsi le κ de Cohen (1960) et ses différents avatars, ou, plus récemment, le α de Krippendorff (2008) .

Pour un état de l'art récent, complet et très fouillé de la question de l'estimation de l'accord interannotateur, on pourra consulter (Mathet, 2017a). Dans cette introduction nous nous contenterons de constater que ces métriques diffèrent uniquement par leur façon de corriger l'accord brut observé sur le corpus par une estimation de la part de chance due au hasard dans cet accord. Les subtilités statistiques sous-jacentes à cette estimation de l'accord par hasard ne sont pas sans intérêt. Il a ainsi été démontré qu'elles avaient un impact sur la qualité d'estimateur de la reproductibilité, en particulier dans le cas d'annotations ordinales (valeurs discrètes ordonnées) (Antoine *et al.*, 2014). Dans cet article, nous souhaitons aborder toutefois une question bien trop négligée de notre point de vue, alors qu'elle revêt une grande importance pratique : il s'agit du manque d'intelligibilité des valeurs d'accord retournées par ces métriques.

Quelle que soit la métrique considérée, celle-ci retourne une valeur d'accord corrigée par la chance d'une valeur maximale de 1 (accord parfait). Du fait de la correction

statistique, il est difficile de relier cette valeur d'accord avec une idée claire de la qualité de l'annotation. Au cours du temps, de nombreux auteurs (cf. figure 1, page 51) ont proposé des seuils de qualité acceptable pour chaque métrique, sans que ces seuils ne reposent sur une argumentation démontrée. Cet article présente précisément des résultats expérimentaux qui tendent à faire sortir l'évaluation de l'accord interannotateur du seul argument d'autorité en matière de seuils d'acceptabilité. Nous proposons pour cela de revenir au critère de reproductibilité qui a fondé les recherches sur le sujet. Notre étude vise, en effet, à relier la mesure de l'accord interannotateur à l'estimation du taux de modifications de l'annotation que l'on obtiendrait avec un autre ensemble d'annotateurs.

Dans un premier temps, nous allons faire une brève présentation de la problématique de l'évaluation de l'accord interannotateur en nous attachant avant tout à décrire tous les facteurs qui peuvent influer sur l'estimation de cet accord, et les difficultés qu'il y a à interpréter les mesures obtenues. Nous proposerons ensuite de réinterpréter ces mesures en reliant la mesure d'accord interannotateur avec la stabilité de l'annotation obtenue avec n'importe quel ensemble d'annotateurs d'une taille donnée. Ainsi, nous lions directement accord interannotateur et reproductibilité de l'annotation.

Cette réinterprétation passe par la mise en place d'une batterie d'expérimentations portant sur des données annotées réelles ou simulées. La quatrième section de cet article présente en détail le cadre méthodologique que nous avons adopté pour cette étude, en insistant en particulier sur la technique de génération d'annotations simulées à partir de données réelles que nous avons utilisée. Nous présenterons enfin les résultats expérimentaux que nous avons obtenus en nous focalisant dans un premier temps sur le κ de Cohen. Ces résultats suggèrent qu'il est possible d'établir une corrélation entre valeurs d'accord interannotateur et taux de reproductibilité de l'annotation. La conclusion nous permettra enfin de détailler l'ensemble des études qui restent à conduire pour arriver à une interprétation directe réellement opérationnelle des métriques d'accord interannotateur.

2. Estimation de l'accord interannotateur : état de l'art et limitations

2.1. *Processus d'annotation*

L'annotation recouvre des processus d'enrichissement de corpus variés, parmi lesquels on distingue deux grandes classes d'activités, qui peuvent, suivant la tâche concernée, correspondre à deux étapes successives d'annotation (Mathet, 2017b) :

– la segmentation, appelée *unitizing* chez Krippendorff (2013), consiste à localiser des unités dignes d'intérêt dans le flux langagier. Elle peut être continue, à savoir qu'elle couvre l'ensemble d'un texte à annoter, ou discontinue. Dans ce second cas, seules quelques portions du texte à annoter seront localisées. C'est par exemple le cas de l'annotation en entités nommées ou en mentions référentielles, où seuls quelques mots ou expressions polylexicales seront localisés dans le continuum du texte ;

– la catégorisation revient, quant à elle, à associer une description linguistique à des unités déjà caractérisées. Elle peut se limiter à associer une catégorie à chaque unité. C'est par exemple le cas d'une annotation en entités nommées où l'on associerait un type d'entité (personne, organisation, lieu...) à chaque unité. La catégorisation peut également être bien plus fine et associer tout un ensemble de traits qualificatifs aux entités étudiées.

La catégorisation peut concerner le document dans sa globalité, auquel cas elle ne succède pas à une étape préalable de segmentation. Dans d'autres situations, la détermination des unités d'intérêt est immédiate, voire automatisable, et ne pose pas de problème de fiabilité. Toutefois, le processus d'annotation englobe en général les deux étapes de segmentation et de catégorisation. La nature très différente de ces deux activités (localiser et qualifier) fait que les guides d'annotation recommandent souvent de les réaliser de manière séparée. Il semble, dès lors, avisé d'évaluer séparément la fiabilité de ces deux opérations. C'est, par exemple, la démarche que nous avons adoptée pour le corpus ANCOR annoté en coréférence (Muzerelle *et al.*, 2014).

Il reste bien sûr envisageable de conduire une évaluation unique, intégrant segmentation et catégorisation, de la fiabilité des annotations. Se pose alors la délicate question de l'alignement des segmentations dans le calcul de l'accord, ainsi que celle de l'estimation d'un accord par chance sur cette délimitation des segments. La famille de métriques γ définie par Yann Mathet (2017, 2017b) constitue la proposition la plus aboutie en la matière. Ce γ ne résout toutefois pas la question de l'intelligibilité des mesures d'accord interannotateur que nous allons étudier dans cet article.

C'est pourquoi nous avons décidé d'étudier cette question, qui n'a jamais été abordée frontalement à notre connaissance, en nous focalisant sur la question de la fiabilité de l'étape de catégorisation, ainsi que sur la métrique la plus répandue pour l'estimer : le κ de Cohen (1960). Ceci, sans ignorer l'existence de propositions alternatives de métriques, qui feront l'objet d'études ultérieures de notre part.

2.2. *La famille de métriques* κ

Dès qu'une annotation de référence est obtenue à partir de plusieurs annotations concurrentes, l'estimation de sa fiabilité repose sur le calcul de l'accord entre les annotateurs. L'accord brut entre les annotateurs ne peut toutefois tenir lieu de bon estimateur de la qualité de la référence, car il n'intègre pas la part d'aléatoire (accord par chance) qui entre dans la mesure finale observée. Cet accord au hasard a pourtant un impact évident sur l'accord brut observé, puisqu'il est *a priori* plus facile d'obtenir un bon accord lorsque la catégorisation ne concerne que deux classes d'annotation que lorsqu'elle en implique dix. On peut songer au cas limite où il n'y aurait qu'une seule classe d'annotation et où l'accord serait, de fait, parfait dès le départ.

Les mesures de fiabilité de l'annotation en catégorisation corrigent l'accord brut observé par une estimation statistique de la part d'accord qui est due à la chance. Les métriques diffèrent par la manière selon laquelle elles estiment cet accord au hasard.

Considérons une tâche d'annotation nominale qui consiste à affecter à chaque entité une catégorie parmi un ensemble de valeurs totalement indépendantes (par exemple, un type d'entité nommée). Le κ de Cohen est estimé par la formule générale [1], où A_o est l'accord brut estimé entre les annotateurs, et A_e est l'estimation de l'accord qui est dû à la chance :

$$\kappa = \frac{A_o - A_e}{1 - A_e} \qquad [1]$$

Pour estimer A_e, Cohen postule que l'accord par hasard dépend uniquement du comportement individuel de chaque utilisateur, qu'il résume par la distribution statistique des catégories d'annotation utilisées par chacun d'entre eux. Selon cette hypothèse, l'accord entre deux annotateurs sera d'autant plus élevé que leur fréquence d'utilisation de chaque catégorie est proche. Ainsi, pour une tâche de catégorisation avec N items annotés, réalisée par deux annotateurs, A_e est estimé par la formule suivante, où n_c^i correspond au nombre de fois où la catégorie c a été utilisée par l'annotateur i :

$$A_e = \frac{1}{N^2} \cdot \sum_c n_c^1 n_c^2 \qquad [2]$$

Cette estimation correspond à une annotation en catégories nominales par deux utilisateurs. Cohen (1968) a proposé une généralisation de la métrique à une catégorisation ordinale ou multivaluée. Cette situation survient, par exemple, pour l'annotation en émotion, où des tours de parole sont catégorisés suivant une échelle ordinale de valences (– 2 = très négatif, – 1 = négatif, 0 = neutre, 1 = positif, 2 = très positif). Dans ce cas, le κ est dit pondéré, puisque la métrique tient compte du fait qu'un désaccord entre deux catégories proches est moins grave que celui entre deux catégories éloignées. Cette généralisation consiste donc à passer d'une distance binaire entre catégories à une distance euclidienne.

Enfin, Davies et Fleiss (1982) ont, de leur côté, défini une généralisation du κ binaire à un ensemble quelconque d'annotateurs. Dans leur proposition, la valeur de A_e correspond à la moyenne, sur l'ensemble des paires des P annotateurs, des valeurs de A_e définies par la formule [2] pour deux annotateurs.

$$A_e = \sum_c \frac{2}{P(P-1)} \sum_{m=1}^{P-1} \sum_{n=m+1}^{P} \frac{n_c^m n_c^n}{N^2} \qquad [3]$$

Comme le rappellent Artstein et Poesio (2008) dans un état de l'art très complet sur l'accord interannotateur, il n'existe toutefois pas à ce jour de proposition concernant la métrique κ qui soit à la fois pondérée et adaptée à un nombre variable d'utilisateurs. La métrique α définie par Krippendorff (2004) permet au contraire une généralisation qui englobe à la fois un nombre quelconque d'annotateurs et une distance euclidienne entre classes d'annotation. Dans le cadre de cette étude, qui relève avant

tout de la preuve de concept, nous avons considéré la version binaire multi-utilisateur de κ telle que proposée par Davies et Fleiss (1982) (cf. formule [3]). Nos travaux futurs concerneront toutefois aussi bien le α que le κ.

2.3. *Biais d'estimation de l'accord interannotateur*

La question de la pertinence des valeurs fournies par les différentes métriques d'accord interannotateur a fait l'objet de nombreuses études expérimentales ou théoriques. Celles-ci se sont avant tout intéressées aux facteurs d'impact qui peuvent biaiser l'estimation de l'accord interannotateur, indépendamment de tout questionnement sur l'interprétation directe d'une valeur d'accord donnée. On peut citer ici quelques-uns des biais potentiels les mieux identifiés dans la littérature :

1) Biais annotateur et nombre d'annotateurs

Le biais annotateur est une question centrale en termes de fiabilité des données, et il sera au centre des expérimentations présentées dans cet article. Il est en effet directement relié à la notion de reproductibilité, puisqu'il décrit l'influence du comportement idiosyncratique d'un annotateur donné sur la construction de la référence. Il peut être estimé par une mesure (bias index) qui quantifie l'amplitude des variations, entre chaque annotateur, de la distribution de la fréquence d'utilisation de chaque catégorie d'annotation (Sim et Wright, 2005). Le κ cherche précisément à intégrer ce biais dans l'estimation de l'accord par chance. La bibliographie est toujours partagée sur la pertinence de cette prise en compte. S'appuyant sur des considérations purement théoriques, Feinstein et Cicchetti (1990) et Di Eugenio et Glass (2004) affirment que ce biais peut avoir un impact sur les valeurs d'accord obtenues. Ils notent en particulier que l'estimation de l'accord par chance A_e sera biaisé si la distribution des annotations varie fortement d'un expert à l'autre. Artstein et Poesio (2008) réfutent au contraire cet argument, en estimant que cette influence concernera A_o et A_e de concert, ce qui en limite l'impact. En dépit de ces controverses, les études expérimentales menées sur le sujet permettent d'arriver à un consensus sur un point : plus le nombre d'annotateurs mobilisés pour construire la référence est élevé, plus le biais annotateur est limité.

2) Prévalence d'une catégorie donnée

La prévalence traduit l'existence d'une surreprésentation d'une catégorie donnée dans les choix d'annotation des annotateurs, que cette prédominance soit due à un biais d'annotation ou résulte de la nature même des données. Dans une telle situation, la probabilité d'arriver à un accord est bien entendu plus importante : la correction due à l'accord par chance A_e est alors susceptible d'être plus importante, et de réduire d'autant la valeur du κ final (Brennan et Silman, 1992 ; Di Eugenio et Glass, 2004). Sim et Wright (2005) ont proposé là encore de définir une mesure (*prevalence index*) pour quantifier l'importance de la prévalence dans les données annotées. Ce *prevalence index* est directement intégré dans le calcul du PABAK, une adaptation du κ de Cohen cherchant à limiter ce biais (Byrt *et al.*, 1993).

3) Nombre de catégories

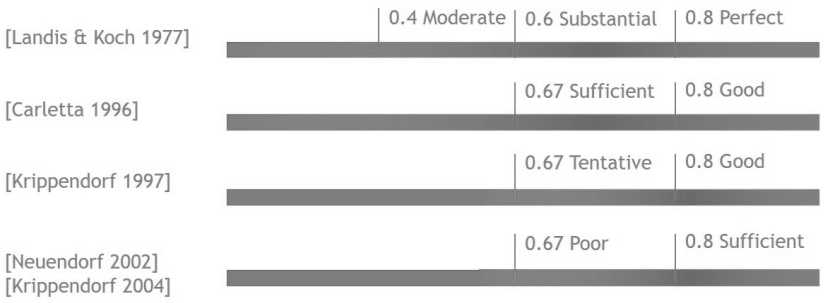

Figure 1. *Échelles subjectives de fiabilité de l'annotation en fonction de l'accord*

Nous avons vu plus haut que le nombre de catégories d'annotation pouvait avoir une influence directe sur la valeur d'accord interannotateur brut Ao. La correction par l'accord au hasard A_e doit limiter cet impact. Les études expérimentales menées sur le sujet ont toutefois montré que les valeurs moyennes de κ observées baissent lorsque le nombre de catégories d'annotation augmente (Brenner et Kliebsch, 1996).

On le voit, les métriques d'accord interannotateur telles que le κ sont potentiellement affectées par de nombreux facteurs d'influence propres aux caractéristiques de l'annotation (nombre de catégories, nombre d'annotateurs, etc.). Ces biais, largement étudiés dans la littérature, rendent difficile l'interprétation d'une valeur d'accord interannotateur donnée. Une même valeur de κ obtenue sur deux annotations différant totalement d'un point de vue méthodologique permet-elle la même conclusion quant à la fiabilité des données annotées, et, dès lors, à partir de quel seuil de κ peut-on estimer qu'une annotation de référence est de qualité suffisante ?

2.4. *Quelle interprétation objective des mesures d'accord interannotateur ?*

L'étude des différents biais qui peuvent entacher la mesure de l'accord interannotateur explique la difficulté qu'il y a à définir une échelle d'interprétation objective de ces mesures. Pourtant, nous avons besoin d'une telle lecture objective, afin de pouvoir associer directement valeur d'accord et fiabilité de l'annotation. Aucune recherche n'a pourtant cherché à creuser cette question à notre connaissance. Ainsi, alors que les métriques d'accord interannotateur mobilisent des calculs statistiques subtils pour nous renseigner sur la qualité de nos données, l'interprétation finale de leur valeur n'a, jusqu'ici, répondu qu'à des échelles subjectives relevant de l'intime opinion de leurs auteurs. On ne compte ainsi plus le nombre de propositions de seuils de bonne fiabilité liés à ces métriques, comme le montre la figure 1.

Tous les auteurs semblent considérer qu'un accord supérieur à 0,8 est un gage relativement satisfaisant de fiabilité des données. Ce seuil porte toutefois une signification très variable d'un auteur à l'autre : là où Landis et Koch (1977) parlent de qualité par-

faite, celle-ci n'est que suffisante par Neuendorf (2002) et Krippendorff (2004). Pour des valeurs inférieures d'accord, les divergences de vue sont encore plus sensibles. À la suite de l'article fondateur de Carletta (1996), le seuil de 0,67 est le plus souvent retenu comme gage de fiabilité acceptable par notre communauté. Elle est pourtant jugée faible par Neuendorf (2002) et Krippendorf (2004), tandis que Landis et Koch (1977) se satisfont même d'un accord interannotateur à 0,4... Cette diversité d'avis résume parfaitement la fragilité méthodologique de ces échelles, qui ne semblent relever que de l'argument d'autorité. Elle est encore évidente lorsque l'on observe qu'un auteur aussi rigoureux que Krippendorff révise lui-même au fil du temps son jugement, sans justifier cette réévaluation. L'objectif des travaux que nous présentons dans cet article est précisément de répondre à cette faiblesse méthodologique, en conduisant des expériences pour élaborer une interprétation objective des valeurs d'accord interannotateur en termes de niveau de reproductibilité.

3. Interpréter les valeurs d'accord en termes de niveau de reproductibilité

L'idée centrale de nos travaux est de revenir à la notion première de fiabilité d'une annotation : une annotation doit être jugée de bonne qualité non pas parce que son accord interannotateur est jugé acceptable suivant une échelle très subjective, mais parce que la référence construite avec tout autre ensemble d'annotateurs reste stable (exigence de reproductibilité). Nous proposons donc d'interpréter la valeur de l'accord interannotateur en termes de taux moyen de variation de la référence que l'on obtiendrait avec d'autres ensembles d'annotateurs.

Notre objectif est donc de conduire une étude expérimentale sur des jeux de données variés, correspondant à des annotations réelles ou simulées à partir de données réelles, pour une correspondance entre valeur d'accord interannotateur et taux de variation de la référence. Nous avons privilégié ici une démarche expérimentale, et non pas une étude statistique théorique, car la question de l'accord interannotateur est hautement multifactorielle (nombreux biais d'estimation bien établis dans la littérature, propositions multiples de métriques d'accord reposant sur des hypothèses fortes sur le comportement des annotateurs qui rendent difficile une théorisation du lien entre accord et stabilité de la référence). Notre article se veut toutefois également une incitation à des tentatives de formalisation plus poussées de ce questionnement méthodologique.

Nous avons observé dans la section précédente que les valeurs d'accord mesurées avec le κ étaient susceptibles de dépendre du nombre d'annotateurs et du nombre de catégories d'annotation. Les correspondances que nous désirons établir à terme dépendront donc de :

– la métrique d'estimation de l'accord (ici : κ) ;

– le nombre d'annotateurs : P ;

– le nombre de catégories : C.

Notre objectif est ainsi d'aller vers une interprétation des mesures d'accord en termes de niveau de stabilité de l'annotation de référence. Ceci pour que les concepteurs de corpus soient à même de juger si le nombre d'annotateurs, de classes ou encore si la complexité de la tâche et la précision du guide d'annotation permettent de fournir des données fiables car reproductibles. Par exemple, un concepteur de corpus qui désirerait s'assurer que son annotation de référence est fiable avec une marge de 5 % d'erreur cherchera la valeur d'accord interannotateur qui lui assure, pour P et C donnés, d'un taux de variation moyen de la référence de 5 % au maximum.

Il convient de noter que l'utilisation d'un vote majoritaire et, entre autres, l'usage d'un tirage aléatoire en cas de ballottage, complexifie encore un peu plus le lien théorique entre valeurs de κ et stabilité de la référence produite. Les travaux qui sont présentés dans cet article sont donc purement expérimentaux et ne fournissent pas encore de tables complètes de fiabilité, mais ils ambitionnent de démontrer la faisabilité et l'intérêt de cette démarche. Nos résultats nous permettront toutefois de proposer des premiers éléments de compréhension du lien entre stabilité de la référence et échelles de fiabilité subjectives proposées dans la littérature. Notons enfin que nos travaux s'inscrivent dans un champ d'application bien précis : celui d'une annotation de type catégorisation d'observables, obtenue par vote majoritaire sur les annotations individuelles et réalisée par plusieurs (au moins trois) experts.

4. Méthodologie expérimentale

Cette partie décrit l'ensemble du cadre expérimental qui a été développé pour estimer le taux moyen de variation d'une annotation en regard de l'accord interannotateur observé sur une annotation de référence. Dans un premier temps, nous décrivons les principes sous-jacents à l'estimation de la stabilité d'une annotation, puis les données, réelles et ensuite simulées, qui ont été utilisées pour nos expériences.

4.1. *Estimation de la stabilité d'une annotation : principes*

Considérons une annotation de type catégorisation réalisée par une population de P annotateurs dont la tâche est d'associer à chacune des N entités d'intérêt (ou *observables*) du corpus une catégorie parmi un ensemble de C catégories. L'annotation est considérée comme parfaitement reproductible si n'importe quel groupe de k personnes, choisies au hasard dans la population, produit toujours la même annotation. Cette annotation idéale correspond à celle que produirait un nombre infini d'annotateurs. Si la reproductibilité n'est pas parfaite, plusieurs annotations différentes peuvent être observées suivant le groupe considéré.

Soit $\mathcal{G} = \{G_i | i \in \{1 : n\}\}$, un ensemble de n groupes comportant chacun k annotateurs, pour des valeurs de k et n données. On note κ_i l'accord interannotateur du groupe G_i. Chacun des groupes G_i produit une référence $R_i = (r_{i_j}, j \in \{1 : N\})$ (par exemple par vote majoritaire), l'annotation la plus fréquente sur l'ensemble

des annotateurs de tous les groupes rassemblés étant considérée comme la *référence absolue*, notée $R = (r_j, j \in \{1 : N\})$. On note τ_i, le taux d'erreurs de la référence produite par le groupe G_i par rapport à la référence absolue R. τ_i se définit par :

$$\tau_i = \frac{1}{N} \sum_{j=1}^{N} d(r_{i_j}, r_j) \quad [4]$$

où $d(a, b) = 0$ si $a = b$ et $d(a, b) = 1$ si $a \neq b$ (distance discrète).

L'accord interannotateur moyen calculé sur l'ensemble $\mathcal{G}$ est la moyenne des accords κ_i calculés sur chacun des groupes G_i :

$$\kappa_{\mathcal{G}} = \frac{1}{n} \sum_{i=1}^{n} \kappa_i \quad [5]$$

alors que le taux d'erreurs moyen calculé sur $\mathcal{G}$ se calcule par

$$\tau_{\mathcal{G}} = \frac{1}{n} \sum_{i=1}^{n} \tau_i \quad [6]$$

Si le nombre d'annotateurs P est suffisamment grand pour que l'on puisse considérer que la référence absolue obtenue est proche d'une référence idéale, et si k est suffisamment petit par rapport à P, alors $\tau_{\mathcal{G}}$ peut être considéré comme une approximation valide du taux de variation attendu de la référence construite avec k annotateurs avec une référence idéale.

Compte tenu des données dont nous disposons, à savoir un corpus annoté par P annotateurs, nous réalisons cette approximation en respectant la procédure suivante :

1) nous calculons la référence absolue R suivant une procédure de vote majoritaire entre les votes des P annotateurs ;

2) nous considérons un nombre k d'annotateurs, avec $k < P$. Pour que la notion de vote majoritaire entre ces annotateurs devienne efficiente, k doit être strictement supérieur à 2[1]. D'où : $2 < k < P$;

3) les sous-ensembles de k annotateurs parmi P sont au nombre de $\complement_P^k$. Suivant leur nombre, ils sont considérés en totalité ou non[2] pour obtenir un ensemble $\mathcal{G}$ de n groupes de k annotateurs avec n assez grand (au moins plusieurs centaines) ;

4) pour chaque sous-ensemble G_i de k annotateurs, nous calculons leur accord interannotateur κ_i et construisons la référence R_i produite par leurs annotations ;

5) nous construisons la référence absolue R correspondant aux P annotateurs par vote majoritaire. Nous calculons également un accord interannotateur théorique $\kappa_{\mathcal{G}}$ comme étant la moyenne des κ_i obtenus sur les différents groupes de k annotateurs ;

1. Lorsque deux classes obtiennent le même nombre de votes sur un observable donné, elles sont départagées par un tirage aléatoire.

2. Pour limiter les temps de calcul, nous nous limitons au tirage aléatoire de 1 000 d'entre eux lorsque leur nombre dépasse plusieurs milliers.

6) la comparaison entre la référence $R = (r_j, j \in \{1 : N\})$ et les références $R_i = (r_{i_j}, j \in \{1 : N\})$ nous donne alors le taux de variation moyen obtenu à partir des annotations produites par les groupes de k annotateurs (cf. formules [4] et [6]) :

$$\tau_{\mathcal{G}} = \frac{1}{n} \sum_{i=1}^{i=n} \sum_{j=1}^{j=N} d(r_j, r_{i_j}) \qquad [7]$$

où d est la distance discrète.

À l'issue de cette procédure, on dispose d'une valeur d'accord interannotateur $\kappa_{\mathcal{G}}$ qui peut être mise en correspondance avec une moyenne des taux de variation $\tau_{\mathcal{G}}$.[3]

4.2. *Corpus et annotations*

Nos expérimentations ont été menées sur les données annotées de quatre corpus différents[4], chacun d'entre eux ayant été collecté pour une tâche spécifique et pour les besoins d'autres projets dans lesquels certains d'entre nous étaient partenaires. Nous sommes en présence d'une simple annotation en classes, tout problème de segmentation préalablement résolu. Ces corpus ont été également choisis car ils impliquaient un grand nombre d'annotateurs. Cette caractéristique sert nos objectifs expérimentaux, mais nous verrons que nos conclusions concernent également les annotations à nombre réduit de codeurs (supérieur ou égal à trois).

4.2.1. *Corpus en émotion*

Une annotation en émotion consiste à ajouter au texte des informations quant à l'émotion que peut ressentir son lecteur. Il n'existe pas de consensus concernant la façon de décrire une émotion dans une tâche d'annotation mais, quelle que soit l'approche choisie, les accords entre les annotateurs sont généralement médiocres, voire faibles, ce qui conduit à prendre les avis d'un grand nombre de codeurs pour définir une annotation de référence (Schuller *et al.*, 2009).

Nos travaux concernaient la détection automatique de l'émotion dans des contes destinés aux enfants. Considérant la difficulté de la tâche, nous avons choisi l'approche

3. Une autre procédure a été expérimentée sur les données réelles : elle consiste à comparer les références produites par deux groupes disjoints de k annotateurs pris parmi les P annotateurs disponibles, ce qui a nécessité d'imposer sur k la contrainte : $2 < k < \frac{P}{2}$. Moyenné sur toutes les paires de groupes disjoints de k annotateurs possibles, le calcul du taux d'erreurs donne également une approximation de la non-reproductibilité. Cette procédure est celle qui avait été mise en œuvre dans Antoine *et al.* (2014). Elle donne des résultats très proches de la procédure utilisée dans les travaux relatés dans cet article.

4. Ces corpus sont disponibles, à fin de reproductibilité, directement auprès des auteurs, à savoir : jean-yves.antoine@univ-tours.fr, jeanne.villaneau@univ-ubs.fr ou encore anais.halftermeyer@univ-orleans.fr.

la plus simple, qui consiste à classifier les émotions suivant une échelle multidimensionnelle ; les deux dimensions essentielles sont la *valence* qui permet de préciser si l'émotion est positive, négative ou neutre et l'*intensité* qui précise le niveau de l'émotion ressentie. Une annotation ordinale en cinq classes {– 2, – 1, 0, 1, 2} permet de réaliser une classification qui combine la *valence* et l'*intensité* réunies. Cette annotation se réduit à trois classes {– 1, 0, 1} si seule la *valence* est considérée.

Le corpus émotion utilisé dans cette étude regroupe 230 phrases issues de deux textes peu connus (Vassallo, 2004 ; Vanderheyden, 1995) et annotées par 25 codeurs. Pour les besoins de nos travaux, deux annotations différentes ont été réalisées : chaque phrase considérée isolément (annotation hors contexte) ou présentée dans l'ordre du récit (annotation en contexte)[5].

4.2.2. *Corpus d'opinion*

Les principes de l'annotation en opinion sont très similaires à ceux pratiqués pour l'opinion : la *polarité* y joue le rôle de la *valence* et le terme d'*activation* peut être utilisé pour désigner l'intensité.

Le corpus est composé de 183 phrases qui expriment des avis déposés sur le site `www.allocine.fr`. Elles ont été annotées par le même groupe d'annotateurs, avec la même échelle de valeurs et dans des conditions strictement semblables à celles du corpus précédent, y compris pour ce qui est de la présentation des phrases hors et en contexte.

4.2.3. *Corpus de coréférence*

Le corpus d'annotation en coréférence utilisé ici est un échantillon produit pour mesurer l'accord interannotateur sur une ressource d'envergure, le corpus ANCOR (Muzerelle *et al.*, 2014). Cet échantillon, qui correspond à un dialogue court à forte interaction a été annoté par 9 annotateurs experts (étudiants de master, doctorants, et enseignants-chercheurs dans la thématique). La tâche de résolution de la coréférence peut être découpée en trois phases :

1) l'identification des mentions référentielles (segmentation) ;

2) l'identification de lien de coréférence entre paires de mentions (segmentation) ;

3) le typage de la relation de coréférence pour chaque paire dont une relation a été identifiée (catégorisation).

Nous utilisons ici uniquement la dernière sous-tâche (3), chaque annotateur a dû choisir pour chaque paire proposée une classe de relation parmi cinq disponibles :

– coréférence directe : les deux mentions présentent la même tête lexicale (*la maison ... cette maison*) ;

5. Pour plus de précisions sur le corpus et son annotation, on pourra se reporter à (Le Tallec *et al.*, 2011).

– coréférence indirecte : les deux mentions présentent deux têtes lexicales différents (*la maison ... cette demeure*) ;

– coréférence pronominale : la seconde mention est un pronom (*la maison ... elle*) ;

– anaphore associative : les deux mentions ne sont pas coréférentes mais il est nécessaire de connaître l'interprétation sémantique de la première pour comprendre la seconde (*cette maison ... la porte*) ;

– anaphore associative pronominale : cette relation présente la même dépendance sémantique entre mentions que pour l'anaphore associative et la seconde mention est un pronom (*cette maison ... ils* s'agissant des habitants de la maison).

4.2.4. *Similarité entre phrases*

Évaluer la similarité entre deux phrases est une tâche classique du TAL, souvent utilisée comme sous-tâche de tâches plus ambitieuses, telles que le résumé automatique. SemEval a proposé des confrontations de systèmes sur ce thème à partir de 2012 et jusqu'en 2017, pour l'essentiel en langue anglaise.

Les données utilisées dans cette étude correspondent à deux petits corpus en langue française créés pour les besoins de nos travaux : le premier a pour thème la conquête spatiale, le second porte sur le thème des épidémies. Pour chacun des corpus, soixante-dix phrases ont été sélectionnées. Dix d'entre elles ont servi de phrases de référence, qui contiennent des informations importantes sur le domaine testé. Chacune de ces phrases a été associée à six autres phrases du corpus, choisies pour que différents niveaux de similarité avec elle soient représentés. Dans chacun des corpus, les soixante paires de phrases ont été annotées par dix annotateurs[6].

L'annotation en similarité est une tâche largement différente de celle en émotion ou opinion, ainsi que de celle qui concerne la coréférence : elle nous semblait donc importante pour augmenter la généricité de notre étude. Pour adopter les données à notre expérimentation, nous avons défini des seuils dans l'échelle des annotations permettant un partage des paires de phrases annotées en trois et en cinq classes[7].

4.3. *Données artificielles*

Notre ambition étant d'étudier une correspondance entre toute valeur d'accord interannotateur et la stabilité de la référence suivant le groupe d'annotateurs choisi, il était indispensable de disposer de données liées à différentes valeurs de cet accord, en l'occurrence κ. La seule façon réaliste d'y parvenir est la génération de données fictives. En même temps, pour garder la possibilité d'étudier l'influence de la tâche, il convient, dans ces données artificiellement générées, de préserver ce qui la caractérise, particulièrement la distribution des annotations initiales.

6. Pour plus de précisions, se reporter à (Vu *et al.*, 2015).

7. Dans le cas de trois classes, les intervalles d'annotations sont $[0\ ;\ 1[$, $[1\ ;\ 2,5[$ et $[2,5\ ;\ 4]$; pour cinq classes, les intervalles sont $[0\ ;\ 0,5]$, $[0,6\ ;\ 1,5]$, $[1,6\ ;\ 2,5]$, $[2,6\ ;\ 3,2]$ et $[3,3\ ;\ 4]$.

4.3.1. *Génération d'annotations fictives*

Le problème consiste donc à générer des annotations fictives qui permettent de modifier les accords interannotateurs tout en respectant la répartition des désaccords dans les données réelles. La méthodologie adoptée est la suivante.

Supposons que l'on ait les données réelles de P annotateurs a_i, $i \in \{1 : P\}$ qui aient donné chacun leur avis sur N observables o_j, $j \in \{1 : N\}$.

Pour chaque observable o_j, soit r_j la référence obtenue par vote majoritaire sur cet observable. On pose $e_{ij} = 0$ si l'observateur a_i a voté pour la référence r_j concernant l'observable o_j et $e_{ij} = 1$ sinon. Suivant cette définition, le nombre e_j des « erreurs » commises par les annotateurs sur l'observable o_j s'obtient par $e_j = \sum_{i=1}^{i=P} e_{ij}$ et f_j, la fréquence de ces « erreurs » se définit par $f_j = \frac{e_j}{P}$.

Par ailleurs, on définit Meo le nombre moyen d'erreurs par annotateur :

$$Meo = \frac{\sum_{i=1}^{i=P}\sum_{j=1}^{j=N} e_{ij}}{P} = \frac{\sum_{j=1}^{j=N} e_j}{P} = \sum_{j=1}^{j=N} f_j$$

Pour générer des annotations fictives, on crée des groupes d'annotateurs plus ou moins « bons », et donc on fait varier le paramètre κ en jouant sur le nombre moyen d'erreurs commises par les annotateurs. Par ailleurs, les paramètres f_j permettent de respecter la répartition des désaccords entre les observables. L'utilisation simultanée de ces deux paramétrages est assurée par le protocole défini ci-dessous.

1) On choisit, en fonction du désaccord que l'on veut obtenir entre les k annotateurs que l'on veut créer, un intervalle $[M - A,\ M + A]$ dans lequel on va faire varier le nombre d'erreurs commises par chacun d'entre eux. Dans la pratique, le centre M de l'intervalle doit être inférieur à Meo, si l'on désire améliorer l'accord obtenu dans les annotations réelles, et supérieur dans le cas contraire. Par ailleurs, le choix de l'amplitude A permet de jouer sur la dispersion du nombre d'erreurs entre les différents annotateurs fictifs créés.

2) Pour chaque annotateur fictif créé, on tire au sort le nombre d'erreurs (nbe) qu'il va commettre dans l'intervalle choisi précédemment $[M - A,\ M + A]$: on crée ainsi un annotateur plus ou moins « bon », à l'intérieur des limites que l'on s'est fixées.

3) On effectue un tirage aléatoire des nbe observables pour lesquels l'annotateur va choisir une annotation autre que la référence. Ce tirage est pondéré par les fréquences $f_j,\ j \in \{1...N\}$ des erreurs sur les N observables présentes dans les annotations réelles. On respecte ainsi la répartition des désaccords dans les données initiales.

On obtient donc ainsi un groupe fictif G de k annotateurs, qui ont chacun un nombre d'erreurs nbe compris entre $M - A$ et $M + A$, et dont les erreurs respectives par rapport à la référence se répartissent préférentiellement sur les observables ayant fait l'objet de désaccords dans les annotations réelles.

5. Résultats expérimentaux

5.1. *Validation de l'idée sur données réelles*

La première expérience que nous avons conduite a consisté à étudier sur des annotations réelles la pertinence du principe de correspondance entre la valeur de κ et le taux de variabilité de la référence. Nous avons pour cela travaillé sur les corpus annotés en émotion et en opinion. Partant de ces ressources annotées par 25 personnes, nous avons créé, pour multiplier les observations, trois sous-corpus obtenus par vote majoritaire en conservant à chaque fois $P = 9$ annotations (respectivement les annotations [1, 9], [9, 17] et [17, 25][8]). Nous avons alors appliqué notre méthode d'estimation du taux de variabilité pour plusieurs valeurs de k annotateurs. Dans cette section, nous présentons à titre illustratif les résultats obtenus pour $k = 3$ à partir des corpus annotés avec trois classes d'émotion ou opinion. Des résultats équivalents sont observés avec d'autres valeurs de k et avec cinq classes d'émotion.

Nos observations (cf. tableau de la figure 2) couvrent une large amplitude de valeurs de κ comprises entre 0,25 et 0,66. On observe que ces variations ont un impact sensible sur l'estimation de la probabilité de reproductibilité de l'annotation, qui varie de son côté de 78 % à 93 %. Ce résultat confirme notre intuition, mais surtout, la figure 2 suggère l'existence d'une forte corrélation entre les valeurs d'accord et la probabilité de variabilité. Il semble donc *a priori* possible d'interpréter les valeurs de κ en termes de reproductibilité de l'annotation, ce qui intéresse directement le concepteur d'un corpus annoté. Il est toutefois nécessaire de disposer de jeux de données en plus grand nombre pour confirmer ces premières observations. C'est l'objectif des expérimentations que nous allons présenter, qui ont été réalisées sur les données simulées.

5.2. *Validation de l'idée sur données simulées*

Nous avons reproduit nos expériences en simulant pour chaque corpus 1 000 nouvelles annotations avec notre procédure de génération de données fictives à partir d'annotations réelles. Cette procédure a été renouvelée pour un nombre d'annotateurs variant entre 2 et 8. La figure 3 présente les résultats ainsi obtenus à partir du corpus annoté avec cinq classes d'opinion. Une fois encore, notons que des résultats équivalents ont été obtenus à partir des autres corpus.

On retrouve ici l'existence d'une forte corrélation négative entre les valeurs d'accord et l'estimation de la probabilité de variabilité (et donc une corrélation positive avec la probabilité de reproductibilité), mais validée cette fois sur un nombre très significatif de jeux de données.

8. Nous utilisons deux fois les annotations 9 et 17 pour une raison purement pratique : obtenir trois groupes d'annotateurs à partir de vingt-cinq. Ce découpage n'induit pas de biais sur l'accord puisque nous ne conservons aucune paire d'annotateurs entre chaque groupe d'annotateurs.

Émotion	HC1	HC2	HC3	EC1	EC2	EC3
Taux variabilité	0,22	0,198	0,206	0,217	0,178	0,164
κ_3	0,248	0,323	0,335	0,273	0,359	0,423
Opinion	HC1	HC2	HC3	EC1	EC2	EC3
Taux variabilité	0,119	0,100	0,109	0,099	0,099	0,087
κ_3	0,557	0,598	0,6	0,571	0,611	0,655

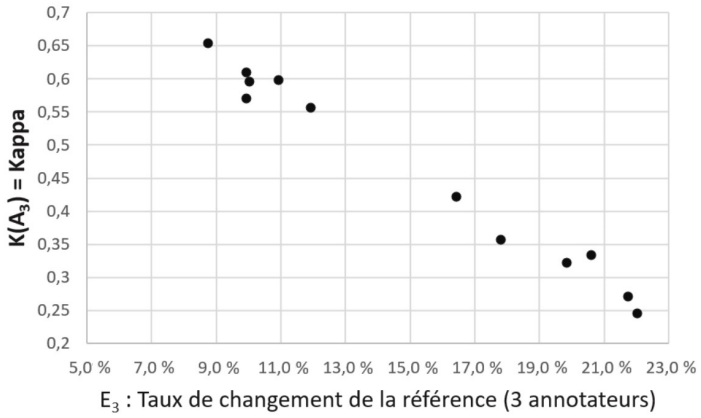

Figure 2. *Résultats sur les corpus en émotion et en opinion à trois classes hors (HC) et en contexte (EC), avec trois annotateurs : valeurs brutes du κ et du taux de modifications de la référence (tableau) et leur représentation graphique*

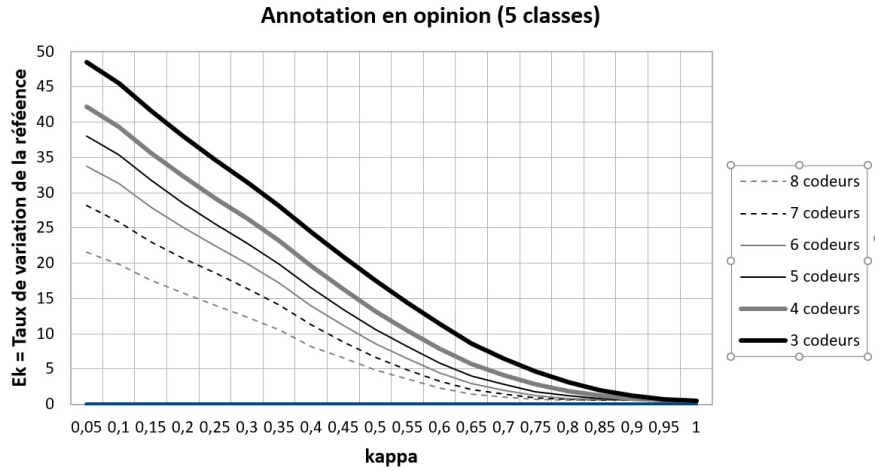

Figure 3. *κ_5 et estimation du taux E_k de variabilité de la référence, pour des données simulées avec k = 2 à 8 annotateurs à partir de l'annotation en cinq classes d'opinion*

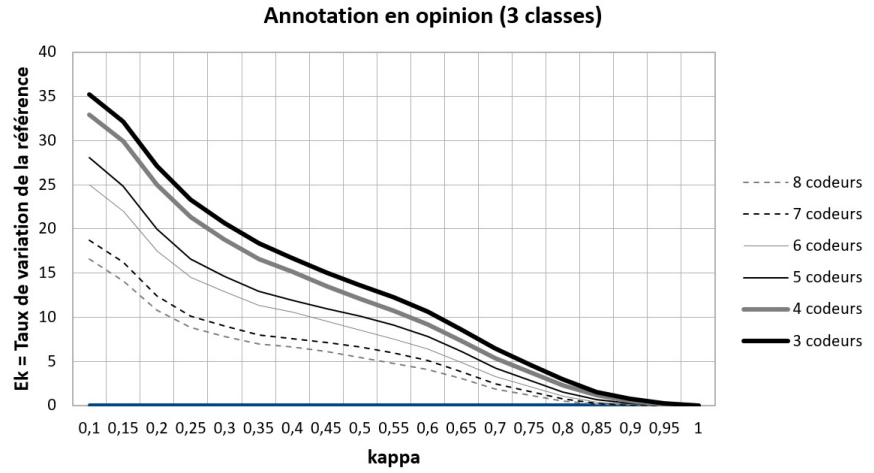

Figure 4. κ_3 *estimation du taux* E_k *de variabilité de la référence, pour des données simulées avec k = 2 à 8 annotateurs à partir de l'annotation en trois classes d'opinion*

On remarque par ailleurs que cette relation de corrélation entre accord et reproductibilité dépend du nombre d'annotateurs considéré. Pour une valeur de κ donnée, la probabilité de variation de la référence est en effet bien plus élevée si celle-ci a été obtenue avec deux annotateurs qu'avec huit. Si l'on considère par exemple un seuil de fiabilité de données de 0,8 pour le κ, la référence est susceptible de changer dans 7 % des cas avec deux annotateurs, alors que cette probabilité est inférieure à 1 % pour huit annotateurs.

Cette influence du nombre d'annotateurs est facilement interprétable. Il est en effet plus difficile de faire bouger une annotation obtenue par vote majoritaire si celui-ci a été obtenu avec un nombre d'annotateurs plus élevé. Comme nous allons le voir, d'autres facteurs peuvent avoir un impact sur le niveau de reproductibilité.

5.3. *Généricité : influence du nombre de catégories*

Nous avons vu que le nombre de catégories d'annotation pouvait influencer les mesures d'accord observées en corpus (Brenner et Kliebsch, 1996). La comparaison des résultats obtenus sur des jeux de données générés à partir d'annotations à cinq catégories et ceux générés sur trois catégories va nous renseigner sur l'influence du nombre de catégories sur nos observations. À titre d'exemple, la figure 4 donne les résultats obtenus à partir du corpus annoté en opinion, mais cette fois avec seulement trois classes.

On note tout d'abord que l'on retrouve ici la même corrélation négative entre les valeurs d'accord et le taux de variabilité, et ce, pour tous les nombres d'annotateurs considérés. En revanche, les courbes présentées sur la figure 4 se caractérisent par des taux de variation de la référence significativement moindres que sur la figure 3, correspondant à une annotation à cinq catégories.

Ce résultat s'explique simplement pour une annotation par vote majoritaire. Considérons un observable annoté donné dans le corpus. Le nombre de votes que reçoit chaque catégorie est *a priori* plus important dans une annotation à nombre réduit de catégories, puisqu'alors les choix d'annotation se distribuent entre un nombre moindre d'annotations. Dès lors, il est plus difficile de faire bouger une majorité portant sur un plus grand nombre de votes, d'où une variabilité plus faible. On en conclut donc que les tables de correspondance que nous souhaitons établir doivent également considérer le nombre de catégories d'annotation.

Nous cherchons d'ailleurs à mieux caractériser cet impact, et voir si une estimation théorique d'une modification de la référence, par hasard, en fonction du nombre de classes d'annotation ne pourrait pas rapprocher les observations. D'un point de vue pratique, il nous semble toutefois important de donner une estimation brute de la reproductibilité d'une annotation aux concepteurs de corpus.

5.4. *Généricité : influence de la tâche*

Les résultats que nous avons présentés jusqu'ici convergent tous vers le fait qu'il semble possible de proposer une interprétation des valeurs de κ sous forme de stabilité de la référence. D'un point de vue pratique, la question se pose toutefois de savoir s'il est possible de définir une échelle unique d'interprétation des mesures d'accord. Nous avons déjà observé qu'une telle échelle ne peut s'entendre que pour un nombre d'annotateurs et de catégories donné. Mais la question la plus importante reste de savoir si la tâche d'annotation donnée influe, elle aussi, sur l'interprétation. Nos corpus d'expérimentations recouvrent une diversité significative de tâches et d'objets linguistiques (émotions, opinions, coréférences, similarités sémantiques). Pour répondre à cette interrogation, nous avons donc décidé de comparer les résultats spécifiques à chaque tâche à un nombre d'annotateurs et de catégories constant, ceci en considérant toujours des corpus simulés à partir des corpus réels.

La figure 5 donne l'ensemble des observations relevées sur nos corpus pour trois annotateurs, et respectivement pour trois et cinq classes d'annotation. On observe une remarquable convergence des courbes entre tous les corpus. Cela suggère la possibilité de définir des intervalles de confiance (en termes de reproductibilité) en fonction du κ qui soient génériques, c'est-à-dire qui ne dépendent pas de la tâche d'annotation considérée.

Ce premier constat doit toutefois être tempéré. Une étude plus attentive montre en effet que les courbes correspondant aux corpus annotés en opinion (courbes pleines) présentent un comportement légèrement différent. Cette différence de comportement

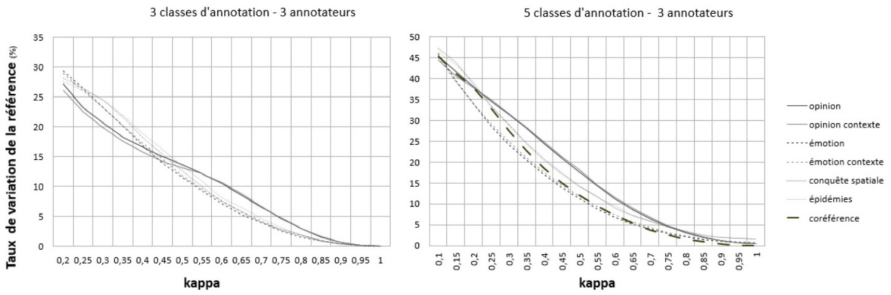

Figure 5. *Comparaison du κ_3 et du taux E_3 de variation de la référence sur tous les corpus, pour des données simulées avec trois annotateurs et trois ou cinq catégories d'annotation*

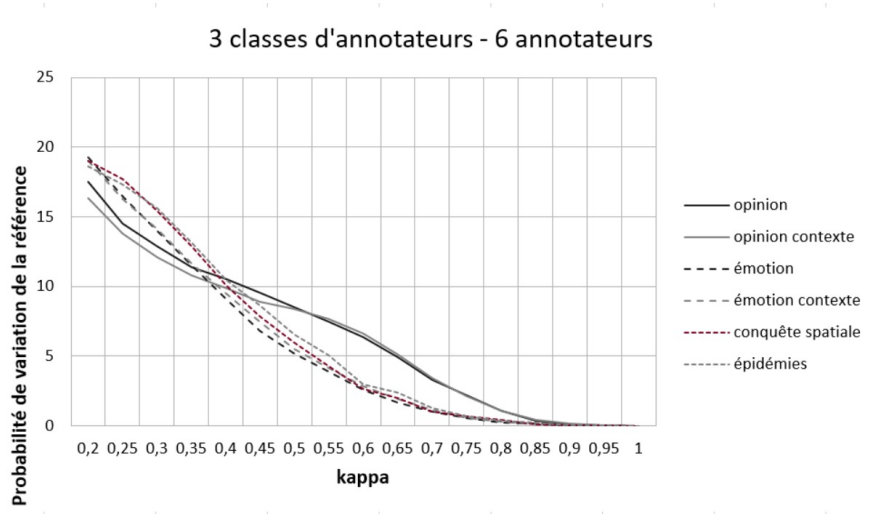

Figure 6. *Comparaison du κ_3 et du taux E_6 de variation de la référence sur tous les corpus, pour des données simulées avec six annotateurs et trois classes d'annotation*

s'accentue lorsque l'on considère un nombre croissant d'annotateurs, comme sur la figure 6. Si les courbes correspondant aux corpus en émotion et en similarité sémantique (épidémie, conquête spatiale) présentent encore une remarquable proximité, la spécificité du corpus en opinion est ici encore plus sensible pour des valeurs de κ intermédiaires : pour les valeurs de κ comprises entre 0,5 et 0,75, la courbe correspondant au corpus en opinion se distingue très sensiblement des autres.

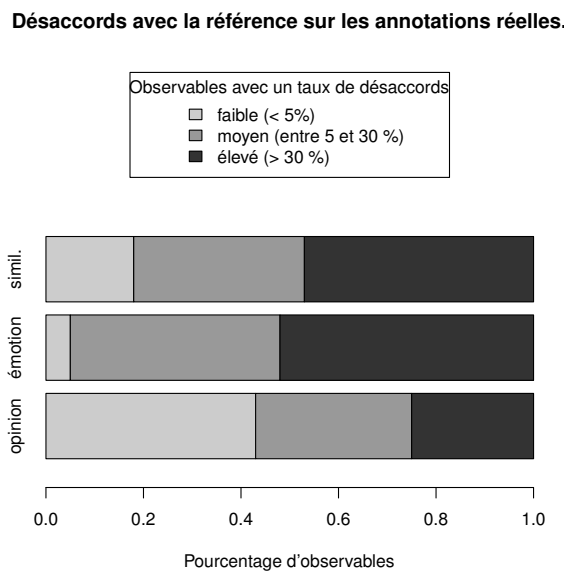

Figure 7. *Histogrammes comparés des désaccords avec la référence sur les corpus émotion, opinion et conquête spatiale*

La nature de l'annotation ne semble pas pouvoir expliquer en elle-même ces divergences de comportement. Comme nous l'avons noté lors de la présentation des corpus (cf. 4.2.2), l'annotation en opinion est en effet proche de l'annotation en émotion dans ses principes, et les deux corpus ont été réalisés dans des conditions identiques, alors que leurs courbes sont différentes. Au contraire, le corpus en émotion donne une courbe proche de celle des annotations en similarité sémantique, alors que ces tâches n'ont rien en commun.

Une analyse des corpus semble suggérer une explication de nature statistique. L'annotation en opinion se caractérise en effet, par rapport aux autres corpus, par une distribution différente des « erreurs » d'annotation par rapport à la référence. Comme on peut le voir dans l'histogramme de la figure 7, le corpus en opinion se caractérise, en effet, par un pourcentage important d'observables (supérieur à 40 %) pour lesquels il y a une quasi–unanimité des annotateurs.

Il reste à élaborer un modèle mathématique qui rendrait compte exactement de l'impact de cette variation de la distribution des divergences d'annotation entre observables. Mais le fait que les autres corpus conduisent à des courbes d'évolution proches, suggère que notre démarche est générique en termes de tâche d'annotation. La section

suivante cherche à éclaircir l'impact de cette distribution non homogène des écarts à la référence.

6. Étude de l'impact de la distribution des divergences d'annotation

Selon nos observations, la distribution des divergences d'annotation entre les observables pourrait être un facteur qui modifie la correspondance entre les valeurs de κ et la stabilité de la référence des différents groupes d'annotateurs. Les résultats présentés dans cette section ont été réalisés sur des annotations entièrement fictives spécialement créées pour confirmer ou infirmer cette observation et pour pouvoir, ultérieurement, explorer d'autres paramètres fréquemment cités, tels que la prévalence.

6.1. *Création des données fictives*

On veut contrôler le nombre de classes (*C*), le nombre d'annotateurs dans chaque groupe créé (*k*), l'accord entre les annotateurs de chaque groupe (le κ), la distribution du taux de désaccords entre les annotations et la référence produite, ainsi que les prévalences entre classes. En l'absence de référence réelle, les taux de variation de la référence seront calculés entre les paires de groupes d'annotateurs créés.

Le protocole adopté est très proche de celui présenté dans la section 4.3.1. Il en diffère cependant sur les points suivants :

– en l'absence de données réelles, une référence initiale des annotations est définie aléatoirement sur les observables[9] ;

– à côté du paramètre M qui gère le nombre d'observables pour lequel un annotateur est en désaccord avec la référence, on introduit un nouveau paramètre : son écart-type σ. Les nombres de désaccords pour chacun des k annotateurs d'un groupe sont tirés aléatoirement suivant une distribution normale de paramètres $(M,\ \sigma)$.

Malgré sa simplicité, ce protocole permet de créer un ensemble de n groupes de k annotateurs ayant un κ relativement stable. Par exemple, on peut constater la faible valeur des écarts-types dans le tableau 1 qui donne la moyenne et l'écart-type des valeurs de κ avec $k = 4$ pour $n = 200$ groupes d'annotateurs créés suivant ce procédé. Par ailleurs, si la référence initiale est un artifice efficace pour constituer des groupes d'annotateurs de κ homogène, elle n'intervient pas dans les calculs qui s'ensuivent.

6.2. *Expérimentations et résultats*

Les expérimentations actuellement réalisées concernent essentiellement la distribution des désaccords sur les observables. Les premiers tests concernant la prévalence

9. Pour contrôler la prévalence, on définit le poids de chaque classe ; bien sûr, la référence initiale est créée en respectant ces poids.

κ :	0,247	0,319	0,398	0,484	0,580	0,684	0,799	0,923
Écart-type κ :	0,029	0,028	0,025	0,022	0,020	0,016	0,015	0,010

Tableau 1. *Moyenne et écart-type des valeurs de κ calculées sur 200 groupes de quatre annotateurs*

ne seront pas présentés : ils semblent indiquer une relation complexe entre prévalence et stabilité de la référence qui demande une étude approfondie.

La figure 8 permet de comparer les résultats obtenus dans le cas où les désaccords entre annotateurs sont répartis uniformément sur l'ensemble des observables, avec celui où 20 % des observables donnent lieu à un accord total, les désaccords étant uniformément répartis sur les 80 % restants. On y observe que la courbe qui correspond à une distribution uniforme est, pour un nombre donné d'annotateurs, en dessous de la courbe correspondante où 20 % des observables font l'unanimité. Il apparaît donc nettement qu'effectivement, la distribution des désaccords entre les observables est un paramètre important de la relation entre la valeur de κ et la stabilité de la référence produite par les annotateurs. De façon prévisible, un report des désaccords sur un plus petit nombre d'observables induit une plus grande instabilité de la référence.

7. Conclusion

Les expérimentations que nous avons présentées dans cet article ont cherché à interpréter les valeurs d'accord interannotateur (ici, le κ de Cohen) sous la forme d'une probabilité de reproductibilité de l'annotation. Cette étude a été menée sur des corpus réels relevant de tâches d'annotation variées, puis a été complétée sur des données d'envergure simulées à partir de ces corpus réels. On peut donc espérer que sa portée est suffisamment large pour nous permettre de tirer certaines conclusions génériques sur la question.

Les résultats que nous avons détaillés montrent qu'il existe une corrélation forte entre la mesure d'accord interannotateur liée à une annotation obtenue par vote majoritaire et la probabilité de reproduire la même annotation de référence avec un autre ensemble d'annotateurs. La nature exacte de cette corrélation reste encore à préciser, une fois mieux modélisée l'influence de facteurs tels que le nombre de classes d'annotation, le nombre de codeurs et surtout la distribution des divergences d'annotation suivant les observables. Il nous semble toutefois que ces travaux sont une piste encourageante vers une interprétation des valeurs d'accord interannotateur utiles au concepteur de corpus annotés. Quelques conclusions prudentes peuvent à ce sujet déjà être tirées de cette étude.

Un κ de 0,8 est une valeur seuil acceptée par tous les auteurs comme gage de bonne fiabilité des données annotées. Nos expériences confirment ces opinions objectives, puisqu'elles se traduisent par une probabilité de variation de l'annotation relativement

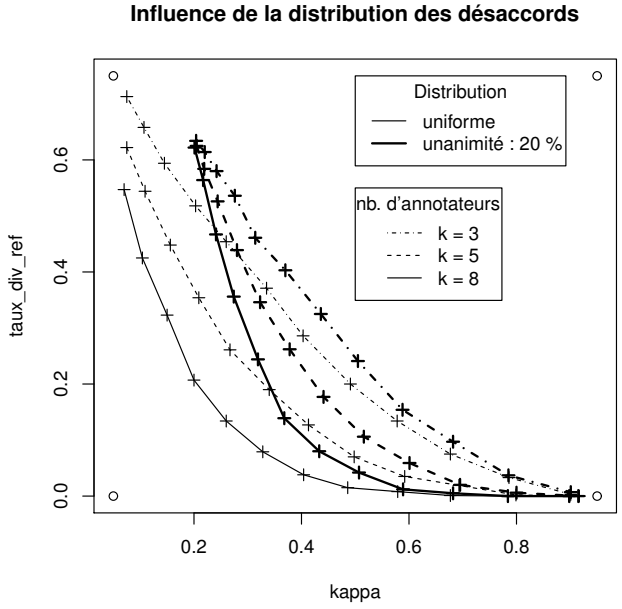

Figure 8. *Résultats comparés : répartition uniforme des désaccords contre 20 % des observables faisant l'unanimité*

faible. Ainsi, toutes nos expériences avec trois annotateurs ou plus montrent que, pour un κ de 0,8, l'annotation de référence a toujours moins de 3 % de chances d'être modifiée avec un autre ensemble d'annotateurs.

À l'opposé, une valeur de κ de 0,67 semble constituer une garantie de fiabilité plus modeste : pour trois annotateurs (figure 5 par exemple) la probabilité de variation de l'annotation de référence est en effet comprise entre 5 % et 10 % cette fois. De tels seuils de fiabilité peuvent déjà nous interroger. Ils posent également question pour ce qui concerne la significativité statistique des campagnes d'évaluation menées en TAL : quel crédit donner aux résultats d'une telle évaluation, si l'annotation de référence qui a servi à l'apprentissage ou au test est susceptible de varier de 10 % avec un autre ensemble d'annotateurs ?

Forts de ces résultats encourageants, nous envisageons d'étendre cette étude en intégrant de nouvelles tâches d'annotation, de nouvelles métriques d'accord telles que le α de Krippendorff. Enfin et surtout, nous aimerions mieux caractériser l'influence de la distribution des annotations de la référence sur les mesures de taux de reproductibilité, afin d'arriver à une interprétation directe et générique des valeurs d'accord.

Nous tenons enfin nos corpus librement à disposition de toute personne qui aimerait reproduire ou compléter cette étude, sur demande auprès des auteurs.

Remerciements

Les auteurs remercient la fédération ICVL (Informatique Centre Val de Loire) pour son soutien à cette recherche, dans le cadre du financement du stage de Dany Brégeon.

8. Bibliographie

Antoine J.-Y., Villaneau J., Lefeuvre A., « Weighted Krippendorff's alpha is a more reliable metrics for multi- coders ordinal annotations : experimental studies on emotion, opinion and coreference annotation. », *EACL 2014*, Gotenborg, Sweden, April, 2014. `http://www.aclweb.org/anthology/E14-1058`.

Artstein R., Poesio M., « Bias decreases in proportion to the number of annotators. », *Proceedings FG-MoL'2005*, Edinburgh, UK, p. 141-150, 2005.

Artstein R., Poesio M., « Inter-coder Agreement for Computational Linguistics », *Computational Linguistics*, vol. 34, n° 4, p. 555-596, December, 2008.

Brennan P., Silman A., « Statistical methods for assessing observer variability in clinical measures. », *BMJ*, vol. 304, p. 1491-1494, 1992.

Brenner H., Kliebsch U., « Dependance of weighted kappa coefficients on the number of categories. », *Epidemiology*, vol. 7, p. 199-202, 1996.

Byrt T., Bishop J., Carlin J., « Bias, prevalence and kappa. », *Journal of Clinical Epidemiology*, vol. 46, p. 423-429, 1993.

Carletta J., « Assessing agreement on classification tasks : the Kappa statistic. », *Computational Linguistics*, vol. 22, n° 2, p. 249-254, 1996.

Cohen J., « A coefficient of agreement for nominal scales. », *Educational and Psychological Measurement*, vol. 20, p. 37-46, 1960.

Cohen J., « Weighted kappa : nominal scale agreement with provision for scaled disagreement of partial credit. », *Psychological Bulletin*, vol. 70, p. 213-220, 1968.

Davies M., Fleiss J., « Measuring agreement for multinomial data. », *Biometrics*, vol. 38, p. 1047-1051, 1982.

Di Eugenio B., Glass M., « The Kappa Statistic : A Second Look », *Computational Linguistics*, vol. 30, n° 1, p. 95-101, 2004.

Feinstein A., Cicchetti D., « High agreement but low Kappa : the problem of two paradoxes. », *Journal of Clinical Epidemiology*, vol. 43, p. 543-549, 1990.

Krippendorff K., « Reliability in content analysis : Some common misconceptions and recommendations. », *Human Communication Research*, vol. 30, n° 3, p. 411–433, 2004.

Krippendorff K., *The content analysis reader.*, Sage, Beverly Hills, CA, chapter Testing the reliability of content analysis data : what is involved and why., 2008.

Krippendorff K., *Content Analysis : An Introduction to Its Methodology*, Sage, Beverly Hills, CA, chapter 11, 2013.

Landis J., Koch G., « The measurement of observer agreement for categorical data. », *Biometrics*, vol. 33, p. 159-174, 1977.

Le Tallec M., Villaneau J., Antoine J.-Y., Duhaut D., « Affective Interaction with a Companion Robot for vulnerable Children : a Linguistically based Model for Emotion Detection », *Proceedings of LTC'2011, Language Technology Conference*, Poznan, Poland, p. 445-450, 2011.

Mathet Y., A contribution to Computational Linguistics and Natural Language Processing : From the Semantics of Space and Time to Annotations and Agreement Measures, Habilitation à diriger des recherches, Université de Caen Normandie, 2017a.

Mathet Y., « The Agreement Measure γcat a Complement to γ Focused on Categorization of a Continuum. », *Computational Linguistics*, vol. 43, n° 3, p. 0661-0681, September, 2017b.

Muzerelle J., Lefeuvre A., Schang E., Antoine J.-Y., Pelletier A., Maurel D., Eshkol I., Villaneau J., « ANCOR_Centre, a Large Free Spoken French Coreference Corpus : description of the Resource and Reliability Measures », *in* ELRA (ed.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reyjavik, Iceland, May, 2014. `https://hal.archives-ouvertes.fr/hal-01075679`.

Neuendorf K., *The content analysis guidebook.*, Sage, Thousand Oaks, CA, 2002.

Schuller B., Steidl S., Batliner A., « The Interspeech'2009 emotion challenge. », *Proceedings Interspeech'2009*, Brighton, UK, p. 312-315, 2009.

Sim J., Wright C., « The Kappa Statistic in Reliability Studies : Use, Interpretation, and Sample Size Requirements. », *Physical Therapy*, vol. 85, n° 3, p. 257-268, 2005.

Vanderheyden K., « Le Noel des animaux de la montagne. », , `http://www.momes.net/histoiresillustrees/contesdemontagne/noelanimaux.html`, 1995.

Vassallo R.-M., *Comment le Grand Nord découvrit l'été.*, Flammarion, Paris, France, 2004.

Vu H.-H., Villaneau J., Saïd F., Marteau P.-F., « Mesurer la similarité entre phrases grâce à Wikipédia en utilisant une indexation aléatoire », *TALN 2015*, Caen, France, 2015.

Conversion et améliorations de corpus du français annotés en Universal Dependencies

Bruno Guillaume* — **Marie-Catherine de Marneffe**** — **Guy Perrier***

* *Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France*
** *The Ohio State University, Columbus, Ohio, USA*

bruno.guillaume@loria.fr, mcdm@ling.osu.edu, guy.perrier@loria.fr

RÉSUMÉ. *Cet article décrit l'effort d'amélioration de deux corpus du français annotés en dépendances syntaxiques, qui s'inscrit dans le cadre du projet Universal Dependencies (*UD*) qui vise à élaborer un schéma d'annotation syntaxique permettant d'analyser de façon similaire plusieurs langues différentes. Nous avons cherché à rendre plus conformes au schéma* UD *ces deux corpus du français, et nous avons évalué l'impact des modifications apportées aux corpus sur la conformité avec le schéma* UD *et la cohérence interne de leur annotation.*

ABSTRACT. *This paper describes an effort to improve the consistency of two French corpora annotated with the Universal Dependencies (*UD*) scheme. The Universal Dependencies project aims at building a syntactic dependency scheme which allows similar analyses for several different languages. We improved the annotations of the two French corpora to render them closer to the* UD *scheme, and evaluated the changes done to the corpora in terms of closeness to the* UD *scheme as well as of internal corpus consistency.*

MOTS-CLÉS : *corpus du français, syntaxe en dépendances, Universal Dependencies, correction de corpus.*

KEYWORDS: *French corpora, grammatical dependencies, Universal Dependencies, corpus correction.*

1. Introduction

Le projet Universal Dependencies[1] (UD) a pour but de créer un schéma d'annotation syntaxique qui puisse être utilisé pour un grand nombre de langues différentes (Nivre *et al.*, 2016). Pour que le projet aboutisse pleinement, il importe que les corpus annotés selon le schéma UD soient, d'une part, conformes à ce schéma, et, d'autre part, présentent une annotation cohérente au niveau de chaque corpus, mais également entre corpus, et principalement entre les corpus d'une même famille de langues (Zeman, 2015), et *a fortiori* entre les corpus d'une même langue. Dans cet article, nous décrivons l'effort pour harmoniser deux corpus français existants avec le schéma UD : UD_FRENCH-GSD et UD_FRENCH-SEQUOIA. Disposer de plusieurs corpus annotés de façon cohérente au sein d'une même langue a plusieurs avantages : offrir plus de données qui peuvent être utilisées de la même façon (les méthodes des réseaux neuronaux actuelles nécessitent une quantité de données importante) et faciliter la comparaison entre corpus ; une annotation cohérente entre langues permet également des applications translinguistiques. Nous avons donc choisi d'harmoniser les annotations de nos corpus du français avec le schéma UD, ce qui permet de les intégrer aux efforts d'apprentissage multilingue (Zeman *et al.* 2017, Zeman *et al.* 2018). Un schéma d'annotation qui se veut universel présente également certains inconvénients : les particularités de chaque langue ne peuvent être représentées. Transformer un corpus existant au schéma UD nécessite souvent de perdre de l'information qui existe dans le corpus original. Nous retraçons les décisions d'annotation et les corrections apportées aux corpus depuis leur intégration à UD. Pour minimiser la perte d'information, nous avons opté, pour certaines constructions, pour une analyse divergente du schéma UD, mais le schéma UD strict est chaque fois recouvrable de façon automatique. Nous montrons également, par deux méthodes d'évaluation, que les modifications apportées au corpus UD_FRENCH-GSD ont contribué, globalement, à une amélioration de la cohérence interne au corpus.

Un format d'annotation commun tel que UD permet, en théorie, d'offrir une analyse parallèle aux constructions grammaticales semblables dans différentes langues. L'annotation se fait à plusieurs niveaux : segmentation du texte en tokens, partie du discours et traits morphologiques pour chaque token, et relations syntaxiques entre les tokens. Pour la partie du discours, un jeu de dix-sept étiquettes est fixé, et chaque langue doit préciser comment elle utilise ce jeu, notamment en précisant quels mots sont considérés comme des particules et comment sont distinguées certaines étiquettes (les verbes des auxiliaires, les déterminants des pronoms, etc.). Pour la morphologie, un large éventail de traits est proposé et chaque langue est invitée à spécifier quels traits sont pertinents et comment ils doivent être annotés. Le schéma UD propose un système de trente-sept étiquettes principales pour les relations syntaxiques, communes à toutes les langues, tant celles morphologiquement riches que celles *pro drop*[2]. Le schéma propose également des étiquettes secondaires qui visent à prendre en compte

1. `http://universaldependencies.org`
2. Les langues pour lesquelles certains pronoms peuvent ne pas être réalisés syntaxiquement.

les particularités de certaines langues, et qui peuvent donc varier entre familles de langue, ou d'une langue à l'autre.

Dans UD, l'objectif d'une analyse parallèle a orienté certains choix d'annotation. Pour une même construction grammaticale, quand on passe d'une langue à une autre, les mots lexicaux (noms, verbes, adjectifs et certains adverbes) sont plus stables que les mots grammaticaux (les autres catégories). Pour offrir une analyse parallèle de constructions semblables, les relations syntaxiques portent donc directement sur les mots lexicaux, les mots grammaticaux étant représentés comme marqueurs de ces mots lexicaux. La figure 1 illustre la différence entre les réalisations syntaxiques de prédication non verbale en français et en russe[3]. Les relations en trait continu sont celles entre mots lexicaux ; les relations en pointillé portent sur les mots grammaticaux. Le français utilise une copule, absente en russe. Le français utilise également des déterminants pour exprimer que le nom est défini, alors que le russe utilise une marque morphologique. Le russe utilise également un système de cas, là où le français utilise une préposition. En donnant priorité aux relations entre mots lexicaux, UD offre une analyse parallèle dans les deux langues.

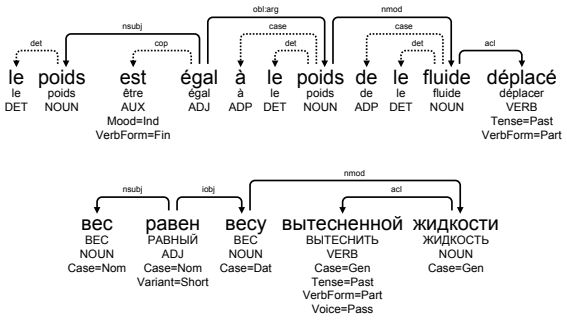

Figure 1. *Annotation en français et russe de la phrase :* le poids est égal au poids du fluide déplacé

Si donner priorité aux relations entre mots lexicaux permet une analyse parallèle entre langues ou différents usages d'une même langue (e.g., langage chez l'enfant qui omet souvent aux premiers stades les mots grammaticaux), certaines particularités linguistiques sont plus difficiles à représenter. On perd en particulier un lien direct entre les verbes et les prépositions qui en dépendent (Osborne et Gerdes, 2019) : par exemple, les prépositions *sur* et *de* dans les phrases verbales *compter sur quelqu'un* ou *dépendre de quelqu'un* seront analysées comme dépendantes de *quelqu'un* alors qu'elles sont sous-catégorisées en fonction du verbe. Un schéma qui permet une analyse parallèle de constructions semblables amène donc à devoir accepter certains compromis d'analyses linguistiques et à des pertes d'information.

3. La phrase russe choisie est dérivée de la phrase `train-s505` du corpus UD_RUSSIAN-GSD.

La version 2.4 de UD contient 146 corpus représentant 83 langues. La taille des corpus diffère toutefois largement, de 292 mots et 55 phrases (UD_TAGALOG-TRG) à environ 1,5 million de mots et 88 000 phrases (UD_CZECH-PDT).

Dans la suite, nous introduisons d'abord les particularités du français dans le projet UD, au niveau de l'annotation ainsi que les corpus disponibles (section 2). Nous décrivons brièvement les méthodes de correction de corpus existantes (section 3). Nous détaillons ensuite les deux corpus qui sont l'objet de cet article, le UD_FRENCH-GSD (section 4) et le UD_FRENCH-SEQUOIA (section 5), et les corrections et transformations que nous y avons apportées pour se conformer au schéma UD. Nous terminons par évaluer l'évolution du corpus UD_FRENCH-GSD en comparant les différentes versions de celui-ci à une annotation de référence ainsi que les performances d'un analyseur syntaxique entraîné et évalué sur les différentes versions.

2. UD pour le français

2.1. *Application du guide d'annotation de UD au français*

Le guide d'annotation général de UD, d'une part, est muet sur certains phénomènes (les clivées par exemple) et, d'autre part, ne traite pas des spécificités de chaque langue. Pour le français, nous avons travaillé avec Marie Candito, Kim Gerdès, Sylvain Kahane et Djamé Seddah à l'établissement d'un guide d'annotation en UD[4] pour les phénomènes ou pour les spécificités du français que ne couvre pas le guide général. Nous présentons ci-après quelques-uns de ces phénomènes et de ces spécificités sur lesquels nous nous sommes penchés plus particulièrement.

2.1.1. *La copule*

Le verbe *être*, lorsqu'il n'est pas auxiliaire de temps ou du passif, est seulement considéré comme copule lorsqu'il a un attribut du sujet qui n'est pas verbal, sinon il est considéré comme un verbe ordinaire. La frontière entre les deux n'est pas toujours évidente. On peut imaginer des critères pour aider à faire la distinction s'inspirant de l'approche pronominale introduite par Claire Blanche-Benveniste (Blanche-Benveniste *et al.*, 1987) et mise en œuvre dans le lexique Dicovalence[5]. Par exemple, si on peut poser la question *est comment ? est quoi ?*, on considérera le verbe *être* comme copule. La première question renvoie à un attribut du sujet qui est une propriété tandis que la seconde renvoie à un attribut du sujet qui est une entité. Si on peut poser la question *est où ?*, le verbe *être* sera analysé comme un verbe ordinaire avec le sens *être situé* et requérant un argument locatif. Ces critères ne permettent pas de trancher tout le temps comme le montrent les exemples (1) et (2). Dans ceux-ci, nous avons choisi de traiter le verbe *être* comme copule mais sans utilisation d'un critère décisif. Les expressions *en poste* et *au pouvoir* sont alors vues comme des propriétés,

4. http://universaldependencies.org/fr

5. https://www.ortolang.fr/market/lexicons/dicovalence

mais on pourrait contester ce choix, en disant que le verbe *être* signifie *être situé* mais dans un sens figuré.

(1) *Il a été en poste de 1934 à 1941.*

(2) *[…] cette ville était encore au pouvoir des Ligueurs.*

Lorsque la copule a comme attribut du sujet une proposition, nous avons choisi de ne pas la considérer comme dépendant de la tête de cette proposition dans une relation `cop`. Si on le faisait, cette tête qui est un verbe aurait deux sujets, un sujet propre et un sujet lié à la copule. Prenons un exemple.

(3) *Le seul problème est qu'il n'a pas de super-pouvoirs […]*

Dans la phrase (3), l'attribut du sujet est toute la proposition *qu'il n'a pas de super-pouvoirs* avec comme tête le verbe *a*. Si, comme on fait généralement dans UD, on considérait qu'il y a une dépendance `cop` de *a* vers *est*, le verbe *a* serait le gouverneur de deux dépendances `nsubj`, l'une pour le sujet propre *il* dans la subordonnée et l'autre pour le sujet de la principale *problème*. Pour éviter cette difficulté, on traite dans ce cas le verbe *être* comme un verbe ordinaire (figure 2).

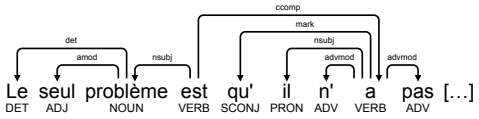

Figure 2. *Annotation* UD *de la phrase (3)*

L'inconvénient est d'avoir deux modélisations différentes de la copule selon la forme de l'attribut du sujet. En anglais et en allemand, par exemple, certains ont choisi de considérer que la copule est dépendante de l'attribut du sujet dans tous les cas mais avec le problème d'avoir deux sujets lorsque l'attribut est une proposition.

2.1.2. *Les dates*

Pour les dates, la hiérarchie suivante dans l'établissement des dépendances a été établie : jour de la semaine → date du jour → mois → année. Dans l'expression *le 25 janvier 2010*, le fait que l'on puisse l'élider en *le 25* est un argument décisif pour que soit choisie comme tête de l'expression *25* et pas *janvier*.

2.1.3. *Les clivées*

Dans une clivée comme (4), la subordonnée n'est pas un argument du foyer de cette clivée, mais pour la distinguer d'une subordonnée ordinaire qui modifie une proposition principale, nous utilisons la relation `advcl:cleft` de la tête du foyer vers la tête de la subordonnée. Une clivée peut avoir une forme interrogative. Dans l'exemple (5),

le foyer de la clivée est le pronom interrogatif *Qu'*. Cette interrogative est une clivée car on peut la paraphraser : *C'est quoi qui va augmenter ?*

(4) *C'est la troisième fois que nous venons, [...]*

(5) *Qu'est-ce qui va augmenter ?*

2.1.4. *Les sujets explétifs*

Le guide d'annotation de UD préconise d'utiliser la relation `expl` pour les arguments syntaxiques qui n'ont pas de rôle sémantique. Récemment, Bouma *et al.* (2018) ont proposé des critères plus précis pour annoter les différentes sous-catégorisations des relations `expl`. Les versions actuelles des corpus présentés sont conformes à ces préconisations pour le pronom réfléchi *se* servant à marquer les verbes essentiellement pronominaux (*s'enfuir*) ou le passif pronominal (*le bruit s'entend de loin*) qui sont bien annotés `expl`.

En revanche, pour d'autres phénomènes, les corpus ne sont pas encore complètement conformes à ces préconisations. Le token *t* dans *va-t-il* est annoté `expl` alors qu'il n'est pas argument. Les annotations actuelles s'écartent également du guide pour les constructions impersonnelles. Dans l'exemple *il ne manque plus que le soleil*, la dépendance de *manque* vers *il* est étiquetée `nsubj`, alors que la dépendance de *manque* vers *soleil* est étiquetée `obj`. Selon le guide d'annotation de UD, la dépendance de *manque* vers *il* devrait être étiquetée `expl` et celle de *manque* vers *soleil* devrait être étiquetée `nsubj`[6]. Ces questions seront gérées dans les prochaines versions du corpus.

2.2. *Les expressions polylexicales*

Pour l'annotation des expressions polylexicales, nous proposons une annotation plus riche[7] que celle prévue dans le guide UD. Cependant, pour fournir des données le plus proche possible de ce qui est préconisé par le guide, nous avons mis en place une conversion automatique. Les données de la distribution officielle sont celles obtenues après conversion (et donc conformes au guide), les données enrichies sont disponibles directement sur le Github du projet (pour le corpus UD_FRENCH-GSD uniquement).

Une définition très générale de ce qu'est une expression polylexicale (EP) consiste à dire que c'est une expression qui ne respecte pas le principe de compositionnalité du sens. Le projet PARSEME-FR[8] propose pour le français des critères permettant d'identifier les EP[9]. Par exemple, le critère [MORPHO] indique que si changer les

6. L'inconvénient d'une telle annotation est qu'elle ne permet pas de distinguer les explétifs sujets des autres explétifs. Par ailleurs, pourquoi annoter une alternance syntaxique pour les constructions impersonnelles et pas pour les constructions passives ?

7. Cette proposition est issue des discussions avec les collègues cités *supra* ainsi qu'une collaboration dans le cadre du projet PARSEME-FR porté par Mathieu Constant.

8. `http://parsemefr.lif.univ-mrs.fr`

9. `https://gitlab.lis-lab.fr/PARSEME-FR/PARSEME-FR-public/wikis/Criteres`

traits morphosyntaxiques d'une expression est agrammatical ou provoque un changement de sens inattendu, elle doit être considérée comme une EP. C'est le cas pour *perdre les pédales* car *perdre la pédale* a bien un sens, mais la différence de sens n'est pas celle attendue pour le changement de nombre d'un objet direct. Comme le reconnaissent les auteurs de ces critères, assigner un statut binaire (EP *versus* expression compositionnelle) est parfois difficile. Leur choix a été de prendre chaque critère comme une condition suffisante d'EP.

Le guide UD impose la relation `fixed` pour représenter les EP : le mot le plus à gauche d'une EP est relié à tous les autres par des dépendances `fixed`. Cette représentation ignore donc la structure syntaxique interne éventuelle d'une EP. Or, quand cette structure est manifeste, il peut être utile de la faire apparaître. Dans notre version enrichie, nous avons choisi de la représenter comme une structure syntaxique ordinaire et nous utilisons des traits pour marquer l'EP. D'une part, la tête de l'EP porte un trait *MWEPOS* qui donne la catégorie grammaticale de l'EP considérée comme un tout par rapport au reste de la phrase. D'autre part, tous les autres composants sont marqués avec le trait *INMWE=Yes*. Nous reprenons ici une proposition de représentation faite par Candito et Constant (2014).

L'exemple (6) contient l'EP *en même temps* et la figure 3 illustre les deux annotations.

(6) *Ils furent créés en même temps que les tribuns de la plèbe.*

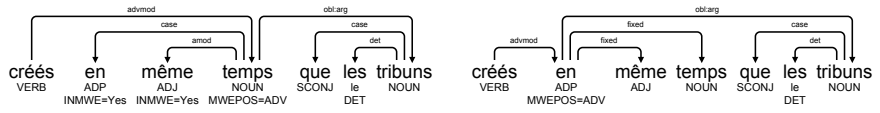

Figure 3. *Annotation enrichie et annotation* UD *pour la phrase (6)*

Dans l'annotation enrichie, comme l'EP joue le rôle d'un adverbe, sa tête *temps* porte un trait *MWEPOS=ADV*. Les composantes de l'EP, *en* et *même*, portent, elles, le trait *INMWE=Yes*. Le caractère facultatif de la conjonction *que* et les coordinations de la forme *en même temps que A et que B* plaident pour dissocier la conjonction *que* de l'EP. Dans l'exemple ci-dessus, la conjonction *que* introduit le complément *que les tribuns de la plèbe* qui est un argument de l'EP, exprimé par une dépendance `obl:arg` de la tête de l'EP *temps* vers *tribuns*.

La seconde annotation de la figure 3, conforme au guide UD est produite automatiquement à l'aide d'un système de réécriture de graphes (sept règles). Il suffit de déplacer la tête de l'EP sur le mot le plus à gauche et de remplacer toutes les dépendances internes à l'EP par des dépendances `fixed` issues du mot le plus à gauche.

Pour les EP qui n'ont pas de structure syntaxique interne évidente (par exemple *ou bien*, *parce que*, *tandis que*), nous utilisons la relation `fixed` prévue par le guide UD. Dans la version 2.4 de UD_FRENCH-GSD, il y a 2 841 EP annotées : 1 274 (45 %) sont annotées en `fixed` et 1 567 (55 %) le sont avec une annotation enrichie.

2.3. *Les corpus du français disponibles au format* UD.

Dans le projet UD (version 2.4 de mai 2019), il y a sept corpus pour le français. Deux de ces corpus (UD_FRENCH-GSD et UD) sont détaillés ailleurs dans l'article. Nous listons les cinq autres ci-dessous. Le tableau 1 reprend la taille des corpus, le nombre de relations grammaticales utilisées, et précise si les corpus disposent de lemmes et de traits morphologiques.

– UD_FRENCH-PARTUT est disponible depuis la version 2.0. Ce corpus est la conversion d'un corpus existant PARTUT (Sanguinetti et Bosco, 2015) et il contient des genres variés (Wikipédia, textes légaux, parole…). Toutes les phrases du corpus sont alignées avec leurs équivalents en anglais (UD_ENGLISH-PARTUT) et en italien (UD_ITALIAN-PARTUT)[10].

– UD_FRENCH-FTB est disponible depuis la version 2.0[11]. La conversion vers le schéma UD a été faite avec les outils de conversion présentés en 5.2 complétés par une étape d'analyse syntaxique automatique pour récupérer les erreurs de conversion par règles (Seddah *et al.*, 2018).

– UD_FRENCH-PUD (disponible depuis la version 2.1) fait partie du projet Parallel Universal Dependencies (PUD) qui propose un ensemble de 1 000 phrases alignées dans différentes langues (Zeman *et al.*, 2017). Les phrases proviennent de Wikipédia ou de textes journalistiques dans cinq langues sources et ont été traduites dans quinze langues au total. Les phrases ont été annotées en suivant le schéma de McDonald *et al.* (2013), et transformées ensuite selon le schéma UD par des membres de la communauté UD. Pour le français, l'annotation de ce corpus diffère quelque peu des autres en restant souvent plus proche de l'anglais : par exemple, les possessifs sont annotés comme des pronoms avec la relation `nmod:poss` (les autres corpus français les annotent comme des déterminants avec la relation `det`).

– UD_FRENCH-SPOKEN (Gerdes et Kahane, 2017), disponible depuis la version 2.2, est une conversion du *treebank* Rhapsodie[12] (MoDyCo *et al.*, 2018) qui contient des annotations en prosodie et en syntaxe de transcriptions de l'oral.

– UD_FRENCH-FQB est disponible depuis la version 2.4. Ce corpus provient d'une conversion automatique du French QuestionBank v1 (Seddah et Candito, 2016) au schéma UD. La conversion a été faite avec la procédure décrite section 5.2.

10. Les corpus italiens et anglais contiennent d'autres phrases absentes de la partie française.
11. Dans les données du projet UD, le texte des phrases et les lemmes ne sont pas fournis car le corpus FTB original est soumis à une licence spécifique (gratuite pour la recherche).
12. `http://www.projet-rhapsodie.fr`

	# phrases	# tokens	lemmes	morpho	# rel
UD_FRENCH-GSD	16 342	400 387	Oui	Oui	50
UD_FRENCH-SEQUOIA	3 099	70 567	Oui	Oui	45
UD_FRENCH-FTB	18 535	573 370	Oui	Oui	36
UD_FRENCH-PARTUT	1 020	28 594	Oui	Oui	45
UD_FRENCH-PUD	1 000	24 734	Non	Oui	42
UD_FRENCH-SPOKEN	2 786	34 972	Oui	Non	52
UD_FRENCH-FQB	2 289	24 135	Oui	Oui	39

Tableau 1. *Information sur les sept corpus du français disponibles en* UD *(version 2.4)*

3. Les méthodes de correction de corpus

3.1. *État de l'art*

Plusieurs méthodes ont été développées pour identifier des erreurs systématiques d'annotation dans les corpus. Il y a deux types de méthodes : celles qui se fondent sur la recherche de motifs définis *a priori* (par exemple De Smedt *et al.* (2016) qui se concentrent sur les erreurs d'annotation dans les expressions figées), et celles qui se fondent sur la localisation de contextes susceptibles de contenir des erreurs (Boyd *et al.*, 2008 ; Alzetta *et al.*, 2018a). Ces deux types de méthodes nécessitent néanmoins une inspection manuelle, pour distinguer les annotations qui résultent d'une réelle différence de celles qui proviennent effectivement d'erreurs.

Boyd *et al.* (2008) se fondent sur le concept de *variation nuclei* développé par Dickinson et Meurers (2003, 2005). Un *variation nucleus* est un élément qui apparaît plusieurs fois dans un corpus avec une annotation différente. Par exemple, dans la figure 4, la construction *ce qui* reçoit deux analyses différentes. Pour Boyd *et al.* (2008), un élément de variation est une paire de mots (*ce* et *élevé* dans l'exemple) apparaissant dans un même contexte (même mot à gauche et à droite du *nucleus*) mais liés par une relation différente.

Figure 4. *Exemple de paire de mots repérée avec la méthode de Boyd* et al. *(2008)*

La méthode de Boyd a été utilisée sur les corpus UD par de Marneffe *et al.* (2017) et Wisniewski (2018) ; de Marneffe *et al.* (2017) l'étendent aux lemmes (et non aux formes de surface), ce qui permet d'extraire plus d'erreurs potentielles (figure 5).

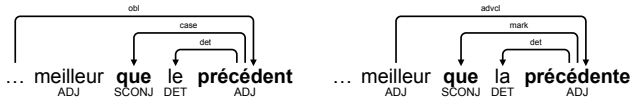

Figure 5. *Exemple d'annotation repérée par la méthode de Marneffe* et al. *(2017)*

Pour UD_FRENCH-GSD 2.0, cette méthode extrait 474 paires potentiellement erronées. Sur 100 paires analysées, 65 % sont erronées.

Alzetta *et al.* (2018b) et Alzetta *et al.* (2018a) utilisent la fiabilité de relations grammaticales produites automatiquement pour identifier des erreurs d'annotation dans les corpus. Leur méthode est fondée sur l'algorithme de Dell'Orletta *et al.* (2013) qui assigne un score de qualité à chaque arc produit par un analyseur, et classe donc les arcs, des plus corrects à ceux potentiellement incorrects. L'algorithme se fonde sur l'hypothèse suivante : plus une structure est fréquemment générée par un analyseur, plus elle est considérée comme correcte. Les relations qui obtiennent un score bas sont en effet souvent susceptibles d'être incorrectes. Une inspection manuelle de ces arcs de pauvre qualité a permis d'identifier des motifs d'erreurs (par exemple, des cas où l'auxiliaire est la tête). Ces motifs d'erreurs, projetés sur tout le corpus italien de UD, reflétaient une réelle erreur dans 86 % des cas.

3.2. *La réécriture de graphes*

La réécriture de graphes est un modèle de calcul puissant qui est utilisé dans de nombreux domaines de l'informatique. Même si les données manipulées dans les modélisations linguistiques sont souvent des arbres, le fait de les voir comme des graphes a de nombreux avantages. D'une part, cela permet de considérer dans un cadre unifié toutes les représentations et notamment celles qui ne sont pas des arbres comme les *enhanced dependencies*[13]. D'autre part, cela permet de gérer de façon homogène des informations périphériques comme l'ordre des mots, par exemple.

Si l'on considère toutes les structures manipulées comme des graphes, la réécriture de graphes est un cadre naturel pour décrire les transformations que l'on souhaite effectuer sur ces structures. Le principe est de décomposer une transformation globale en une succession de transformations élémentaires qui ne font chacune intervenir

13. Les *enhanced dependencies* (`http://universaldependencies.org/workgroups/enhanced.html`) comprennent notamment toutes les dépendances argumentales qui ne sont pas représentées dans UD, telles que les sujets d'infinitifs. On les appelle des dépendances profondes par opposition à celles de UD qui sont considérées comme des dépendances de surface.

qu'un nombre fixé de nœuds du graphe. Chaque transformation élémentaire est décrite par une règle qui est composée de deux parties : un motif (qui décrit le contexte dans lequel une transformation peut s'appliquer) et des commandes (qui décrivent les modifications locales qu'il faut apporter au graphe quand le motif est repéré).

La réécriture de graphes a été utilisée ponctuellement en traitement des langues (Hyvönen, 1984 ; Bohnet et Wanner, 2001 ; Jijkoun et de Rijke, 2007 ; Bedaride et Gardent, 2009 ; Chaumartin et Kahane, 2010 ; Ribeyre, 2016). Nous verrons plus loin des exemples d'utilisation de la réécriture de graphes pour des transformations de corpus, mais on utilise également de façon indépendante le repérage de motifs pour la recherche d'erreurs ou d'incohérences dans les annotations.

Nous utilisons le logiciel GREW[14] pour décrire et appliquer les règles de nos transformations et l'outil dérivé GREW-MATCH[15] pour le repérage de motifs. Nous avons choisi d'utiliser le système GREW, que nous maîtrisons et pouvons faire évoluer selon les besoins, deux des auteurs du présent article étant impliqués dans la création et la maintenance de cet outil (Bonfante *et al.*, 2018).

3.3. *Corrections utilisant une double conversion*

Convertir une annotation d'un format A dans un format B puis convertir l'annotation obtenue en sens inverse de B vers A permet de détecter des incohérences dans l'annotation initiale. Si l'on ne retrouve pas l'annotation de départ après la double conversion, il y a nécessairement une erreur dans l'annotation de départ ou dans le processus de conversion.

Dans les expériences que nous avons menées, nous avons utilisé comme format B le format SUD (*Surface-Syntactic Universal Dependencies*) (Gerdes *et al.*, 2018). Le projet UD se propose de fournir un cadre commun d'annotation syntaxique à des corpus, quelle que soit la langue utilisée par ces corpus. C'est pourquoi le format tient compte du niveau sémantique, qui présente une moindre différence entre les langues. Les têtes des dépendances sont les mots lexicaux, et les mots grammaticaux sont dépendants des mots lexicaux. Les structures syntaxiques de UD sont donc plutôt plates[16]. Par ailleurs, les noms des relations dépendent des fonctions syntaxiques qu'elles représentent, mais aussi des catégories des mots qu'elles font intervenir. Ainsi, une relation *modificateur*, quand le gouverneur est un verbe, est notée `obl`, `advcl` ou `advmod`, selon que le dépendant est un nom, un verbe ou un adverbe. Quand le gouverneur est un nom, la relation est notée `nmod`, `amod`, `acl` ou `advmod` selon que le dépendant est un nom, un adjectif, un verbe ou un adverbe.

14. `http://grew.fr`
15. `http://match.grew.fr`
16. Dans UD_FRENCH-GSD, 4,9 % des phrases ont une annotation non projective contre 14,3 % pour la version SUD.

Le format SUD se veut plus fidèle à la tradition de la syntaxe en dépendances (Mel'čuk, 1988 ; Hajič *et al.*, 2017) qui met au centre les fonctions syntaxiques. Les noms des relations représentent strictement des fonctions syntaxiques indépendamment de la catégorie des mots qu'elles relient (comme également pour le format FTB décrit plus bas). Ainsi, il y a une seule relation `mod` pour les modificateurs qui vaut, quelle que soit la catégorie du mot qui modifie et du mot qui est modifié. Les mots fonctionnels sont la tête des expressions qu'ils introduisent. Cela concerne les prépositions, les conjonctions de subordination, les auxiliaires et les copules. Les seules exceptions concernent les déterminants et les conjonctions de coordination qui sont dépendants de mots auxquels ils s'appliquent.

Comme dans UD, les étiquettes peuvent se présenter en deux parties *(étiquette principale : étiquette secondaire)*, ce qui permet de varier la granularité de l'étiquetage. Par exemple, la relation `comp` s'applique à tous les arguments verbaux mais on peut préciser `comp:obj` pour les objets directs. Le lecteur trouvera plus bas un exemple de phrase annotée en UD et en SUD dans la figure 8.

4. Le corpus UD_FRENCH-GSD

La dernière version en date du corpus UD_FRENCH-GSD est la version 2.4 (novembre 2018). Suivant les conventions du projet UD, le corpus est divisé en trois parties dont les tailles sont reprises dans le tableau 2.

	`train`	`dev`	`test`	Total
Nombre de phrases	14 450	1 476	416	16 342
Nombre de tokens	354 655	35 714	10 018	400 387

Tableau 2. *Taille des parties du corpus* UD_FRENCH-GSD

La maintenance et la correction de corpus sont une tâche de longue haleine que nous menons en continu depuis 2015 sur le corpus UD_FRENCH-GSD dans le contexte du projet UD qui propose deux nouvelles distributions des corpus chaque année.

Les modifications apportées sont de deux types : l'application au corpus de nouvelles décisions d'annotation (qui peuvent être générales au projet UD, spécifiques au français ou internes au corpus lui-même) et la correction d'erreurs ou d'incohérences.

4.1. *Historique du corpus*

Les données actuelles du corpus UD_FRENCH-GSD proviennent du *universal dependency treebank v2.0* de Google (McDonald *et al.*, 2013). Le corpus UD_FRENCH-GSD a été initialement annoté manuellement par deux équipes différentes dans le

cadre du projet de Google d'harmonisation de dépendances grammaticales pour six langues (l'anglais, l'allemand, le français, l'espagnol, le suédois et le coréen). La stratégie adoptée par Google pour harmoniser le schéma d'annotation était de garder l'ensemble des relations grammaticales et les analyses le plus proche possible de celles du corpus anglais SD (de Marneffe *et al.*, 2006). Les annotateurs ne pouvaient ajouter de relations grammaticales que pour des phénomènes non présents en anglais. Cette stratégie rigide d'annotation a, d'une part, poussé à analyser certaines constructions grammaticales parallèlement à l'analyse proposée pour l'anglais au lieu de tenir compte de différences systématiques entre langues et, d'autre part, a gardé des étiquettes différentes pour exprimer une distinction qui est réalisée au niveau morphologique (comme marquer différemment un modificateur infinitif (*un spectacle à mourir d'ennui*) ou participial (*un spectacle barbant au possible*)) et qui ne se manifeste pas dans toutes les langues.

Début 2015, le corpus a été converti dans le format UD sans que cette transformation ne soit documentée. Depuis lors, nous avons travaillé régulièrement pour faire évoluer ces données. Pour la production de la version 2.0, nous avons dû convertir les annotations pour tenir compte des nouvelles conventions d'annotation. Les autres évolutions ont été internes au corpus ou en collaboration avec d'autres corpus du français. Le tableau 3 décrit les principales évolutions du corpus. Le tableau ne fait pas figurer les corrections diverses des annotations et les homogénéisations qui ont été continues pendant les trois ans de travail sur le corpus. Nous revenons dans la section suivante sur ces corrections plus ponctuelles.

Il est important de noter que les données initiales disponibles ne comportaient aucune métadonnée. Pour les phrases du corpus UD_FRENCH-GSD, nous ne disposons ni du texte original ni d'information sur la source dont est issue la phrase. Ainsi, lors de la production de la version 1.2, le champ `sent_id` a été introduit avec une numérotation interne au corpus et le champ `text` a été reconstruit à partir des annotations.

Par ailleurs, dans les données initiales, nous avons trouvé de nombreuses phrases tronquées dans lesquelles une partie du texte original était manquante. La quasi-totalité des phrases tronquées provient de Wikipédia et les parties manquantes sont souvent des unités ou des dates. Nous avons donc décidé de compléter les phrases du corpus en reprenant le texte original de Wikipédia. En effet, ces erreurs sont dues à des traitements précédents d'extraction des données et ne reflètent pas des erreurs d'usage de la langue. En revanche, nous n'avons évidemment pas corrigé d'autres types d'erreurs de grammaire ou de syntaxe qui sont présentes dans la version Wikipédia utilisée pour construire le corpus (même si souvent ces erreurs sont corrigées dans la version actuelle de Wikipédia) car ces erreurs font partie de l'usage de la langue qu'un corpus souhaite refléter. Il arrive encore que l'on trouve de nouvelles phrases tronquées qui sont alors corrigées au fil de l'eau.

Dans l'annotation UD, il n'est pas requis de faire de distinction entre les deux types de compléments obliques du verbe : les arguments et les modificateurs. Cependant, pour les applications et notamment une analyse sémantique, la distinction entre argument et modificateur est essentielle. Nous avons donc décidé de l'ajouter à la res-

source UD_FRENCH-GSD en sous-typant `obl` en `obl:mod`, `obl:arg` ou `obl:agent` (pour les compléments d'agent).

Une petite partie de ces distinctions a pu être prédite automatiquement (par exemple, les compléments antéposés, séparés par une virgule sont systématiquement `obl:mod`). Pour le reste, l'annotation a été faite manuellement par un annotateur à l'aide d'un outil dédié qui présente à l'annotateur les cas à décider, classés par préposition. Ainsi, 39,87 % des `obl` ont pu être sous-typés dans la version 2.2 et 67,40 % dans la version 2.3.

Version	Date	Description
1.0	janvier 2015	Version initiale construite depuis le Google Dataset
1.1	mai 2015	Corrections de quelques segmentations en phrases
1.2	novembre 2015	Ajout de métadonnées, corrections de phrases tronquées
1.3	mai 2016	Ajout des lemmes et de la morphologie
1.4	novembre 2016	Pas de changement majeur
2.0	de mars à mai 2017	Adaptation des données aux nouvelles directives
2.1	novembre 2017	Revue systématique des auxiliaires
2.2	juillet 2018	Sous-typage partiel de la relation `obl`
2.3	novembre 2018	Applications des nouvelles décisions : date, EP (partiel)
2.4	mai 2019	Corrections imposées par le nouveau validateur UD

Tableau 3. *Historique des versions du corpus* UD_FRENCH-GSD

4.2. *Méthodes utilisées pour corriger le corpus*

Une partie des modifications induites par de nouvelles décisions peuvent être faites automatiquement. En revanche, d'autres nécessitent un travail manuel d'annotation. Pour la correction d'erreurs ponctuelles ou d'incohérences, le recours à l'annotation manuelle est systématique.

4.2.1. *Corrections automatiques*

Nous donnons à titre d'exemple une liste (loin d'être exhaustive) des modifications automatiques apportées au corpus UD_FRENCH-GSD.

– Ajout des lemmes et de la morphologie. Dans la version 1.1, pour chaque mot, seule la catégorie était renseignée. Nous avons appliqué systématiquement une prédiction du lemme et de la morphologie à partir de la forme fléchie et de la catégorie en utilisant le lexique *Lefff* (Sagot, 2010). Les annotations ont été faites manuellement pour les mots absents de *Lefff* et pour les lemmes ambigus (par exemple, la forme verbale *suis* peut correspondre aux lemmes *être* ou *suivre*). Pour les ambiguïtés morphologiques, des règles ont été utilisées et les ambiguïtés morphologiques qui n'ont pas pu être levées automatiquement l'ont été manuellement.

– Changement de l'annotation des coordinations pour la version 2 de UD. Dans la version 1, les conjonctions de coordination sont rattachées à la tête du premier conjoint, alors que dans la version 2, elles sont rattachées à la tête du second conjoint.

– Réduction des verbes auxiliaires aux lemmes *être*, *avoir* et *faire* et des copules au lemme *être*. Dans les premières versions, les auxiliaires comprenaient aussi les auxiliaires modaux et d'aspect et dans certaines d'entre elles, les copules étaient étendues à quelques verbes d'état.

– Changement de l'annotation des dates avec le chiffre du jour comme tête.

Dans la figure 6, nous donnons un exemple de règle qui modifie localement l'annotation des dates. La partie *pattern* définit la partie du graphe à modifier. Elle comprend deux nœuds `DAY` et `MONTH` qui représentent le jour et le mois. Le jour est dépendant du mois par une relation `nummod` ou `amod`. La partie *commands* donne la suite des opérations qui va permettre de transformer le motif détecté. La dépendance entre le mois et la date est remplacée par une dépendance `nmod` de `DAY` vers `MONTH`. Les commandes `shift_in` et `shift_out` permettent de modifier la façon dont les nœuds du motif sont reconnectés au contexte. Ici, comme on change la tête de `MONTH` à `DAY`, la commande `shift_in` déplace la dépendance pointant sur l'ancienne tête `MONTH` vers la nouvelle `DAY`; la commande `shift_out` déplace les dépendances issues de l'ancienne tête (sauf celles de type `nmod` et `nummod`).

```
rule day_month {
  pattern {
    DAY [upos=NUM|ADJ]; MONTH [lemma="janvier"|...|"décembre"];
    e: MONTH -[nummod|amod]-> DAY; DAY << MONTH;
  }
  commands {
    del_edge e; add_edge DAY -[nmod]-> MONTH;
    shift_in MONTH ==> DAY; shift_out MONTH =[^nmod|nummod]=> DAY;
  }
}
```

Figure 6. *Une des règles utilisées pour la conversion des dates*

4.2.2. *Corrections manuelles*

Quand les modifications à apporter ne sont pas complètement prédictibles à partir de la syntaxe existante, il faut procéder à une annotation manuelle. C'est aussi le cas pour les corrections ponctuelles : on corrige au cas par cas mais en essayant à chaque fois d'appliquer ces modifications de façon homogène à tout le corpus en recherchant des contextes similaires à celui qu'on est en train de corriger.

Dans de nombreux cas, l'outil GREW-MATCH a été utilisé car il permet justement d'afficher un ensemble de contextes similaires décrits par un motif. Les utilisations de cette méthode ont été très nombreuses (de l'ordre de la centaine) et nous en donnons quelques-unes à titre d'exemple :

– détection des incohérences entre une relation et la nature du dépendant (relation `amod` qui ne pointe pas sur un adjectif par exemple) ;

– recherche de deux relations `sujet` portant sur le même verbe ;

– recherche d'une relation `sujet` dont le gouverneur est un nom sans copule (figure 7) ;

– recherche des défauts d'accord (en genre ou en nombre) entre un verbe et son sujet, un adjectif et le nom sur lequel il porte.

```
pattern { N [upos=NOUN]; V -[nsubj]-> N }
without { V -[cop]-> * }
```

Figure 7. *Motif pour retrouver les noms communs sujets en l'absence de copule*

4.2.3. *Application de la méthode de double conversion*

La double conversion du corpus UD_FRENCH-GSD, telle qu'elle a été présentée à la sous-section 3.3 a été appliquée. Le corpus obtenu (appelé UD_FRENCH-GSD^{DC}) a été comparé avec le corpus initial UD_FRENCH-GSD.

Lors de la première application de la méthode, 3 955 des 400 440 relations (soit 1 %) étaient différentes (gouverneur différent ou étiquette différente). Une différence peut provenir d'une erreur dans l'écriture des règles de conversion qu'il faut alors corriger. Sinon, elle vient d'une erreur d'annotation de UD_FRENCH-GSD. L'erreur est alors corrigée, mais comme il est expliqué plus haut (4.2.2), on vérifie systématiquement les contextes similaires pour corriger de façon globale d'autres erreurs similaires. En itérant la mise à jour des données et de la conversion, le nombre de différences entre UD_FRENCH-GSD et UD_FRENCH-GSD^{DC} a été réduit de 3 955 à 351. Le format UD est plus précis que SUD pour représenter les coordinations imbriquées, ce qui explique 337 différences. Les 14 différences restantes sont liées à des phénomènes très spécifiques non pris en compte par les règles de conversion.

5. Le corpus UD_FRENCH-SEQUOIA

Le corpus UD_FRENCH-SEQUOIA est apparu plus récemment dans le projet UD. Il est disponible depuis la version 2.0 (mars 2017) et nous l'avons obtenu par conversion du corpus SEQUOIA[17] existant. Nous présentons ici le corpus tel qu'il existe au format avant la conversion ainsi que la conversion elle-même.

17. `http://deep-sequoia.inria.fr`

5.1. *Le format d'annotation* FTB

Le corpus SEQUOIA a été annoté en constituants en suivant le schéma d'annotation du *French Treebank* (FTB) (Abeillé et Barrier, 2004) et converti automatiquement en dépendances (Candito et Seddah, 2012b) (nous noterons FTB ce format en dépendances dans la suite). Les dépendances à distance ont été corrigées manuellement (Candito et Seddah, 2012a). Depuis, le corpus a notamment évolué lors de l'annotation avec des dépendances profondes (Candito *et al.*, 2014).

Le format FTB est proche du format SUD en ce sens que les mots fonctionnels sont la tête des expressions qu'ils introduisent, y compris les conjonctions de coordination. En revanche, les auxiliaires sont dépendants des verbes principaux auxquels ils s'appliquent, comme dans UD (figure 8). Alors que dans SUD, le nombre d'étiquettes principales est très réduit, il y en a beaucoup plus dans FTB. Les arguments verbaux sont représentés dans SUD avec la dépendance `comp`, alors que dans FTB, on spécifie `ato` (attribut de l'objet), `ats` (attribut du sujet), `a_obj` (objet indirect introduit par *à*), `de_obj` (objet indirect introduit par *de*), `obj` (objet direct) ou `p_obj.o` (objet indirect introduit par une autre proposition que *à* ou *de*).

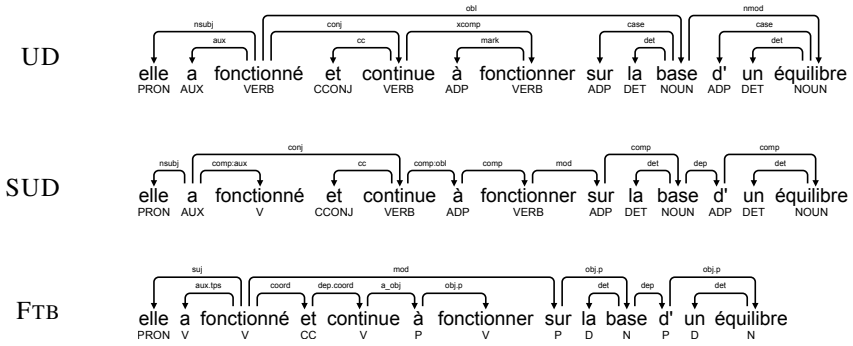

Figure 8. *Annotation de la phrase* elle a fonctionné et continue à fonctionner sur la base d'un équilibre *suivant les trois formats* UD*,* SUD *et* FTB

5.2. *La conversion du corpus* SEQUOIA *dans le format* UD

Les deux formats d'annotation FTB et UD contiennent essentiellement les mêmes informations, exprimées différemment. Il est donc possible d'envisager une conversion automatique de l'un vers l'autre. La plupart des travaux de conversion d'un corpus existant au schéma UD ont utilisé des règles de mapping automatiques, suivies de corrections manuelles (Nivre (2014) pour le corpus suédois Talbanken, Pyysalo *et al.* (2015) pour le corpus finlandais *Turku Dependency Treebank* (TDT), Attardi *et al.* (2016) pour l'*Italian Stanford Dependency Treebank* (ISDT), Taji *et al.* (2017) pour le *Penn Arabic Treebank*, Bouma et Van Noord (2017) pour le corpus néerlandais *Lassy*

Small, Cecchini *et al.* (2018) pour le corpus latin *Index Thomisticus*, Przepiórkowski et Patejuk (2019) pour le corpus polonais POLFIE). Dans certains corpus UD émanant de corpus existants, pour limiter la perte d'information, plusieurs relations secondaires ont été introduites (dans le TDT finlandais (Pyysalo *et al.*, 2015), le corpus polonais SZ (Wróblewska, 2018), le *treebank* hébreu HTB (Sade *et al.*, 2018)). Nous avons toutefois tenté d'éviter cette solution autant que possible.

Sur certains points, le format FTB est plus riche que celui de UD ; il fait par exemple systématiquement la distinction entre modificateur et argument pour les dépendants obliques des verbes, des adjectifs et des adverbes[18]. Cela a permis d'avoir systématiquement le typage `obl:mod` vs `obl:arg` dans UD_FRENCH-SEQUOIA.

Inversement, nous avons été confrontés à quelques cas dans lesquels SEQUOIA manquait d'information pour être correctement converti en UD. Par exemple, les deux exemples *faire installer des poubelles* et *faire jouer les enfants* étaient annotés de façon identique (*poubelles* objet de *installer* et *enfants* objet de *jouer*) alors que dans UD, on a la relation `obj` dans le premier cas et la relation `obj:agent` dans le second. Nous avons décidé d'enrichir l'annotation de départ du corpus SEQUOIA en ajoutant un trait *agent=y* uniquement dans le second cas pour le différencier du premier. Même si cette information est disponible dans l'annotation en syntaxe profonde (Candito *et al.*, 2014), nous avons décidé de ne pas faire la conversion depuis cette version enrichie, notamment pour permettre l'application de cette conversion à d'autres corpus pour lesquels l'annotation en syntaxe profonde ne serait pas disponible.

Comme souvent, lors de la conversion d'un corpus, l'étude des cas qui posent problème amène soit à modifier la conversion elle-même, soit à revoir l'annotation originale.

La conversion est codée sous la forme d'un système de réécriture de graphes qui contient environ 200 règles. Une grande partie de ces règles sont de simples changements d'étiquettes de partie du discours (par exemple, `N` devient `NOUN`) ou de noms des traits morphologiques (par exemple, *g=m* devient *Gender=Masc*). Les autres règles qui modifient réellement la structure en dépendances sont dans l'ordre d'application : le traitement des énumérations et des coordinations (15 règles), les changements de tête (11 règles), le traitement de la copule (1 règle) et les EP (9 règles). La règle de traitement de la copule (nommée `verb_ats`) est donnée dans la figure 9.

```
rule verb_ats {
  pattern {V[upos=VERB, lemma="être"]; e: V-[ats]->A}
  without {A[upos=VERB, VerbForm=Inf|Fin]}
  commands {del_edge e; shift V==>A; add_edge A-[cop]->V; V.upos=AUX}
}
```

Figure 9. *Règle* `verb_ats` *de traitement de la copule*

18. Pour les groupes prépositionnels dépendant de noms, le choix a été fait comme dans UD de ne pas trancher la distinction entre modificateur et argument car elle est trop difficile à annoter.

La clause `pattern` décrit les conditions dans lesquelles cette règle doit s'appliquer : le verbe *être* est la source d'une dépendance `ats`. La clause `without` est une condition négative qui permet de décrire une exception et bloque l'application de la règle si la cible de la dépendance `ats` est un verbe (la justification de cette condition négative est donnée à la sous-section 2.1.1). Si ces conditions sont vérifiées, on applique dans l'ordre les quatre commandes : suppression de l'arc repéré, changement de tête (tous les liens attachés au verbe *être* sont reportés sur l'attribut), ajout de l'arc `cop` en sens inverse et changement de la catégorie de *être* de `VERB` à `AUX`. La figure 10 donne un exemple d'application de la règle.

Figure 10. *Annotation d'une phrase avant et après l'application de la règle* `verb_ats`

Le fait de traiter la copule après les remplacements d'étiquettes permet d'avoir des règles plus simples pour les changements d'étiquettes qui doivent tenir compte de la catégorie du dépendant. Dans la phrase *Je dois être honnête* la relation entre *doit* et *être honnête* est changée en `xcomp` car la tête de *être honnête* est *être* avant l'application de la règle `verb_ats`. Si cette règle avait été appliquée avant, la tête de *être honnête* serait *honnête* et la prédiction de la relation `xcomp` nécessiterait un motif plus complexe.

6. Évaluation

6.1. *Comparaison avec une annotation de référence*

Parmi les méthodes d'évaluation, la comparaison avec une annotation de référence est essentielle car elle permet de confronter les annotations à l'avis d'expert. Malheureusement, cette évaluation est coûteuse en temps et pas toujours facile à réaliser.

Comme nous voulons observer l'évolution de la qualité de l'annotation au fil des versions, nous avons dû construire une référence avec des phrases issues de ce corpus. Nous avons donc extrait aléatoirement 108 phrases (soit 2 502 tokens) du corpus UD_FRENCH-GSD (partie `train`) pour lesquelles une annotation manuelle a été faite par trois annotateurs. Chaque phrase a été annotée par deux des trois auteurs et les différences ont fait l'objet de discussions communes pour construire l'annotation de référence utilisée ensuite pour l'évaluation. Le corpus ainsi obtenu est noté UD_FRENCH-GSD^{GOLD}.

Pour éviter que l'annotation actuelle des 108 phrases dans UD_FRENCH-GSD ne biaise la construction de la référence, les phrases ont été préannotées en utilisant un

	A1/A2	A1/A3	A2/A3	Moyenne
Résultats bruts	89,11	85,97	93,42	89,50
Après normalisation des dates	90,33	88,08	93,42	90,61

Tableau 4. *Accords entre annotateurs pour le corpus de référence (108 phrases)*

Version	2.0	2.1	2.2	2.3	2.4
Exactitude (%)	85,13	87,86	88,35	91,12	92.00

Tableau 5. *Exactitude pour les différentes versions du corpus* UD_FRENCH-GSD

analyseur à base de règles (Guillaume et Perrier, 2015) qui est indépendant du corpus. Inversement, les annotations de ce corpus UD_FRENCH-GSD^{GOLD}n'ont pas été utilisées dans d'autres parties du travail et notamment pour la détection d'erreurs. En revanche, un biais évident de notre évaluation est que les trois annotateurs impliqués dans la création de la référence le sont également dans la mise à jour du corpus.

Les valeurs d'exactitude calculées dans les tableaux 4 et 5 ont été obtenues avec le script `conll17_ud_eval.py` fourni avec la version 2.0[19]. La mesure utilisée est le *Weighted LAS* pour laquelle seul le poids de la ponctuation est mis à zéro.

Le tableau 4 donne les accords entre les annotateurs (%) pour les trois paires d'annotateurs. Une partie significative des différences avec l'annotateur A1 porte sur l'annotation des dates. La dernière ligne du tableau donne les scores d'accord si on ne tient pas compte de ces différences.

Dans le tableau 5 figurent les évaluations des différentes versions du corpus par rapport aux données de UD_FRENCH-GSD^{GOLD}. Les conventions d'annotation ayant changé lors de la version 2.0, il n'est pas pertinent d'évaluer les versions 1.x. La progression de l'exactitude des versions 2.0 à 2.4 par rapport au corpus UD_FRENCH-GSD^{GOLD} est significative et nous pensons que, malgré les biais évoqués plus haut, cela reflète une évolution de la qualité du corpus UD_FRENCH-GSD.

19. `https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2184`

6.2. *Estimation de la cohérence de l'annotation à l'aide d'un analyseur syntaxique*

Une autre méthode permettant d'avoir des indications sur la qualité d'un corpus ou tout du moins sur la cohérence des annotations qui y figurent est d'entraîner un analyseur syntaxique et d'évaluer ses performances. Dans cette expérience, nous avons utilisé l'analyseur UDPIPE (Straka et Straková, 2017) que nous avons entraîné avec le sous-corpus `train`, puis utilisé pour analyser les deux autres sous-corpus disponibles (`dev` et `test`). Nous avons utilisé les paramètres par défaut de UDPIPE. En effet, notre but n'est pas d'optimiser les performances de l'analyseur mais d'évaluer l'évolution du corpus et donc d'utiliser les mêmes paramètres pour chacun de nos tests.

La figure 11 montre l'exactitude obtenue par l'analyseur UDPIPE sur les sous-corpus `dev` à gauche et `test` à droite. De nouveau malgré les biais inhérents à ce type d'évaluation, nous observons une progression significative qui montre que la cohérence interne du corpus a nettement progressé.

Ces évaluations montrent également une différence très nette entre les niveaux d'exactitude obtenus sur le sous-corpus `dev` d'une part et sur le sous-corpus `test` d'autre part. Cette différence était déjà manifeste dans les données de la version 1.1 (5,79 % de différence pour le LAS) et elle est moindre mais toujours présente dans la version 2.4 (2,13 %). Nous ne savons pas expliquer cette différence car nous n'avons pas d'information sur la façon dont les différents sous-corpus ont été construits dans les données originales. En revanche, dans nos modifications, nous avons toujours traité l'ensemble du corpus sans tenir compte de ce découpage.

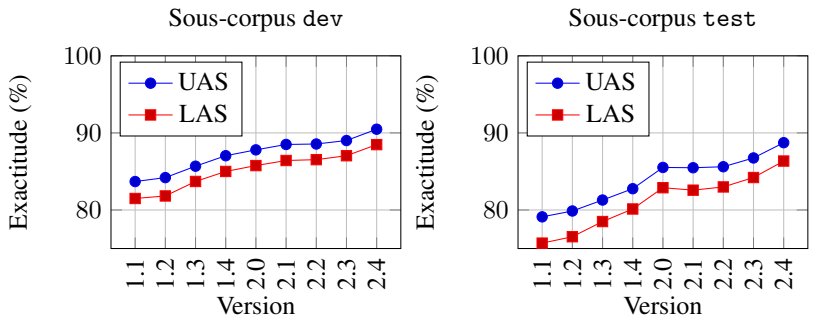

Figure 11. *Performance de* UDPIPE *entraîné puis évalué sur les différentes versions de UD_French-GSD*

Les mêmes expériences menées sur le corpus UD_FRENCH-SEQUOIA ne montrent pas d'évolution significative (les annotations ont très peu changé entre les versions 2.0 à 2.4). Toutes les valeurs de UAS sont entre 87,27 et 88,05 et celles de LAS entre 85,61 et 86,59. Ces valeurs sont inférieures de 1 à 3 points par rapport à celles obtenues pour les versions récentes de UD_FRENCH-GSD, sans doute car le

corpus d'apprentissage est nettement plus petit (50 536 tokens pour UD_FRENCH-SEQUOIA contre 354 655 tokens pour UD_FRENCH-GSD).

7. Conclusion

Dans cet article, nous avons décrit le travail effectué pour faire évoluer des corpus existants. Les deux corpus ont fait l'objet de démarches assez différentes. Pour UD_FRENCH-GSD, nous avons travaillé uniquement sur les données de la version 1.1 de UD et nous ne disposions pas des données annotées initialement avant conversion ni des métadonnées. Pour UD_FRENCH-SEQUOIA, en revanche, nous avons essentiellement travaillé sur la question de la conversion du format initial du corpus vers le format UD. Pour évaluer la qualité d'une annotation, il est courant d'avoir recours à une annotation par un expert sur une petite partie des données qui servent alors de référence. Nous avons appliqué cette méthode au corpus UD_FRENCH-GSD. Sur le même corpus, nous avons également observé la cohérence des annotations en entraînant un analyseur syntaxique sur une partie de données et en l'appliquant au reste. Ces deux évaluations montrent une progression de la cohérence et de la qualité du corpus.

Le contexte du projet UD permet aujourd'hui un travail beaucoup plus systématique pour harmoniser les données de différents corpus et les rendre plus faciles à comparer. Même si l'on reste dans un cadre monolingue, comme c'est le cas dans cet article, l'harmonisation entre les corpus développés par différentes équipes reste une question compliquée et qui nécessite beaucoup de concertation. Nous avons largement entamé cette harmonisation en collaboration avec les collègues travaillant sur d'autres corpus du français, mais il reste encore beaucoup à faire pour améliorer la situation pour les prochaines versions. Une suite naturelle à ce travail serait également d'élargir cette harmonisation vers d'autres langues et au moins, dans un premier temps, aux autres corpus en langue romane. Une autre direction pour enrichir les données décrites ici sera d'annoter dans les corpus du français les *Enhanced Universal Dependencies*[20] qui facilitent l'utilisation des corpus pour l'analyse sémantique de la langue.

Remerciements

Les auteurs remercient Alane Shur, Matias Grioni et Carly Dickerson qui ont participé à l'annotation de UD_FRENCH-GSD ; ils remercient également l'un des relecteurs et le comité de la revue dont les commentaires détaillés et pertinents ont permis d'améliorer le document. Ce travail a bénéficié de l'infrastructure du CPER LCHN (Contrat Plan État Région « Langues, Connaissances & Humanités Numériques »), ainsi que d'un *Google Faculty Research Award* attribué à Marie-Catherine de Marneffe.

20. `http://universaldependencies.org/u/overview/enhanced-syntax.html`

8. Bibliographie

Abeillé A., Barrier N., « Enriching a French Treebank. », *Proceedings of LREC 2004*, 2004.

Alzetta C., Dell'Orletta F., Montemagni S., Simi M., Venturi G., « Assessing the Impact of Incremental Error Detection and Correction. A Case Study on the Italian Universal Dependency Treebank », *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, p. 1-7, 2018a.

Alzetta C., Dell'Orletta F., Montemagni S., Venturi G., « Dangerous Relations in Dependency Treebanks », *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16)*, p. 201-210, 2018b.

Attardi G., Saletti S., Simi M., « Evolution of Italian Treebank and Dependency Parsing towards Universal Dependencies », *Proceedings of the Second Italian Conference on Computational Linguistics*, p. 25-30, 2016.

Bedaride P., Gardent C., « Semantic Normalisation : a Framework and an Experiment », *8th International Conference on Computational Semantics - IWCS 2009*, Tilburg, Netherlands, January, 2009.

Blanche-Benveniste C., Deulofeu J., Stéfanini J., Van Den Eynde K., *Pronom et syntaxe : l'approche pronominale et son application au français*, vol. 1, Peeters Publishers, 1987.

Bohnet B., Wanner L., « On using a parallel graph rewriting formalism in generation », *EWNLG '01 : Proceedings of the 8th European workshop on Natural Language Generation*, Association for Computational Linguistics, p. 1-11, 2001.

Bonfante G., Guillaume B., Perrier G., *Application of Graph Rewriting to Natural Language Processing*, John Wiley & Sons, 2018.

Bouma G., Hajic J., Haug D., Nivre J., Solberg P. E., Øvrelid L., « Expletives in Universal Dependency Treebanks », *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, Association for Computational Linguistics, p. 18-26, 2018.

Bouma G., Van Noord G., « Increasing Return on Annotation Investment : The Automatic Construction of a Universal Dependency Treebank for Dutch », *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, Association for Computational Linguistics, p. 19-26, 2017.

Boyd A., Dickinson M., Meurers W. D., « On detecting errors in dependency treebanks », *Research on Language & Computation*, vol. 6, n° 2, p. 113-137, 2008.

Candito M., Constant M., « Strategies for Contiguous Multiword Expression Analysis and Dependency Parsing », *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Association for Computational Linguistics, Baltimore, Maryland, p. 743-753, June, 2014.

Candito M., Perrier G., Guillaume B., Ribeyre C., Fort K., Seddah D., Villemonte De La Clergerie É., « Deep Syntax Annotation of the Sequoia French Treebank », *Proceedings of LREC 2014*, Reykjavik, Iceland, may, 2014.

Candito M., Seddah D., « Effectively long-distance dependencies in French : annotation and parsing evaluation », *TLT 11-The 11th International Workshop on Treebanks and Linguistic Theories*, 2012a.

Candito M., Seddah D., « Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. », *TALN 2012*, Grenoble, France, 2012b.

Cecchini F. M., Passarotti M., Marongiu P., Zeman D., « Challenges in Converting the Index Thomisticus Treebank into Universal Dependencies », *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, ACL, p. 27-36, 2018.

Chaumartin F.-R., Kahane S., « Une approche paresseuse de l'analyse sémantique ou comment construire une interface syntaxe-sémantique à partir d'exemples », *TALN 2010, Montreal, Canada*, 2010.

de Marneffe M.-C., Grioni M., Kanerva J., Ginter F., « Assessing the annotation consistency of the universal dependencies corpora », *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, p. 108-115, 2017.

de Marneffe M.-C., MacCartney B., Manning C. D., « Generating Typed Dependency Parses from Phrase Structure Parses », *Proceedings of LREC 2006*, 2006.

De Smedt K., Rosén V., Meurer P., « Studying Consistency in UD Treebanks with INESS-Search », *Fourteenth Workshop on Treebanks and Linguistic Theories (TLT14)*, 2016.

Dell'Orletta F., Giulia V., Montemagni S., « Linguistically-driven Selection of Correct Arcs for Dependency Parsing », *Computacion y Sistemas*, vol. 2, p. 125-136, 2013.

Dickinson M., Meurers W. D., « Detecting Inconsistencies in Treebanks », *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT-03)*, 2003.

Dickinson M., Meurers W. D., « Detecting Errors in Discontinuous Structural Annotation », *Proceedings of the 43rd Annual Meeting of the ACL*, p. 322-329, 2005.

Gerdes K., Guillaume B., Kahane S., Perrier G., « SUD or Surface-Syntactic Universal Dependencies : An annotation scheme near-isomorphic to UD », *Universal Dependencies Workshop 2018 (UDW 2018)*, Bruxelles, Belgique, November, 2018.

Gerdes K., Kahane S., « Trois schémas d'annotation syntaxique en dépendance pour un même corpus de français oral : le cas de la macrosyntaxe », *Atelier ACor4French, Actes de la 24e conférence sur le traitement automatique des langues (TALN)*, Orléans, p. 1-9, 2017.

Guillaume B., Perrier G., « Dependency Parsing with Graph Rewriting », *IWPT 2015, 14th International Conference on Parsing Technologies*, 14th International Conference on Parsing Technologies - Proceedings of the Conference, Bilbao, Spain, p. 30-39, 2015.

Hajič J., Hajičová E., Mikulová M., Mírovskỳ J., « Prague Dependency Treebank », *Handbook of Linguistic Annotation*, Springer, p. 555-594, 2017.

Hyvönen E., « Semantic Parsing as Graph Language Transformation - a Multidimensional Approach to Parsing Highly Inflectional Languages », *COLING*, p. 517-520, 1984.

Jijkoun V., de Rijke M., « Learning to Transform Linguistic Graphs », *Second Workshop on TextGraphs : Graph-Based Algorithms for Natural Language Processing, Rochester, NY, USA*, 2007.

McDonald R. T., Nivre J., Quirmbach-Brundage Y., Goldberg Y., Das D., Ganchev K., Hall K. B., Petrov S., Zhang H., Täckström O., Bedini C., Castelló N. B., Lee J., « Universal Dependency Annotation for Multilingual Parsing. », *ACL (2)*, ACL, p. 92-97, 2013.

Mel'čuk I., *Dependency Syntax : Theory and Practice*, Albany, N.Y. : The SUNY Press, 1988.

MoDyCo, LaTTiCe, CLLE-ERSS, LPL, IRCAM, « TREEBANK RHAPSODIE », , `https://hdl.handle.net/11403/rhapsodie/v1`, 2018.

Nivre J., « Universal Dependencies for Swedish », *Proceedings of the Swedish Language Technology Conference (SLTC)*, 2014.

Nivre J., de Marneffe M.-C., Ginter F., Goldberg Y., Hajic J., Manning C. D., McDonald R., Petrov S., Pyysalo S., Silveira N. *et al.*, « Universal dependencies v1 : A multilingual treebank collection », *Proceedings of LREC 2016*, p. 1659-1666, 2016.

Osborne T., Gerdes K., « The status of function words in dependency grammar : A critique of Universal Dependencies (UD) », *Glossa*, vol. 4, n° 1, p. 1-28, 2019.

Przepiórkowski A., Patejuk A., « From Lexical Functional Grammar to enhanced Universal Dependencies », *Language Resources and Evaluation*, Feb, 2019.

Pyysalo S., Kanerva J., Missilä A., Laippala V., Ginter F., « Universal Dependencies for Finnish », *Proceedings of NODALIDA 2015*, p. 163-172, 2015.

Ribeyre C., Méthodes d'analyse supervisée pour l'interface syntaxe-sémantique, PhD thesis, Université Paris Diderot, 2016.

Sade S., Seker A., Tsarfaty R., « The Hebrew Universal Dependency Treebank : Past Present and Future », *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, Association for Computational Linguistics, p. 133-143, 2018.

Sagot B., « The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French », *Proceedings of LREC 2010*, La Valette, Malte, mai, 2010.

Sanguinetti M., Bosco C., *PartTUT : The Turin University Parallel Treebank*, Springer International Publishing, Cham, p. 51-69, 2015.

Seddah D., Candito M., « Hard time parsing questions : Building a QuestionBank for French », *Proceedings of LREC 2016*, 2016.

Seddah D., Villemonte De La Clergerie É., Sagot B., Martinez Alonso H., Candito M., « Cheating a Parser to Death : Data-driven Cross-Treebank Annotation Transfer », *Proceedings of LREC 2018*, Miyazaki, Japan, mai, 2018.

Straka M., Straková J., « Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe », *Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, Association for Computational Linguistics, Vancouver, Canada, p. 88-99, August, 2017.

Taji D., Habash N., Zeman D., « Universal Dependencies for Arabic », *Proceedings of the Third Arabic Natural Language Processing Workshop*, ACL, p. 166-176, 2017.

Wisniewski G., « Errator : a Tool to Help Detect Annotation Errors in the Universal Dependencies Project », *Proceedings of LREC 2018*, p. 4489-4493, 2018.

Wróblewska A., « Extended and Enhanced Polish Dependency Bank in Universal Dependencies Format », *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, Association for Computational Linguistics, p. 173-182, 2018.

Zeman D., « Slavic Languages in Universal Dependencies », *Proceedings of Slovko 2015 : Natural Language Processing, Corpus Linguistics, E-learning*, 2015.

Zeman D., Hajič J., Popel M., Potthast M., Straka M., Ginter F., Nivre J., Petrov S., « CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies », *Proceedings of the CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, Association for Computational Linguistics, p. 1-21, 2018.

Zeman D., Popel M., Straka M., Hajic J., Nivre J., Ginter F., *et al.*, « CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies », *Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, Association for Computational Linguistics, Vancouver, Canada, p. 1-19, August, 2017.

Notes de lecture

Rubrique préparée par Denis Maurel

Université de Tours, LIFAT (Laboratoire d'informatique fondamentale et appliquée)

Atefeh FARZINDAR, Diana INKPEN. Natural Language Processing for Social Media, Second Edition. *Morgan & Claypool publishers*. 2017. 175 pages. ISBN 978-1-68173-614-3.

Lu par **Pascal VAILLANT**

Université Paris 13 / LIMICS – UMR INSERM 1142

Atefeh Farzindar et Diana Inkpen, spécialistes de traitement automatique des langues et de sciences des données, ont coécrit cet ouvrage en langue anglaise sur le traitement des contenus textuels des médias sociaux. Le livre consacre une grande partie de ses pages aux nouvelles méthodes que l'on doit concevoir pour adapter les techniques existantes de traitement automatique des langues aux matériaux spécifiques de ce domaine en pleine expansion. Il aborde ensuite les différents domaines d'application de l'analyse textuelle des médias sociaux avec leurs problématiques spécifiques. Enfin, il traite des différentes questions que posent la collecte et l'annotation de ce type de données.

Un volume ahurissant de données est créé chaque jour sur les médias sociaux : en juin 2019, à chaque minute qui s'écoule, trois millions de messages sont postés sur Facebook à travers le monde, et un demi-million sur Twitter. Cette sphère de données est en phase d'expansion rapide.

Une partie importante de l'information de ces messages est constituée de texte en langue naturelle (pas nécessairement la plus importante en termes de quantité d'octets, mais certainement la plus importante en termes de clés de compréhension du contenu des messages). Ces flux de données gigantesques sont porteurs d'informations sur la manière dont les foules réagissent en temps réel aux stimuli de leur environnement et aux nouvelles qui leur parviennent. Chaque utilisateur individuel, en envoyant son message, même insignifiant, contribue plus ou moins consciemment à un vaste hypertexte qui fournit des informations économiques, écologiques, sociologiques et politiques.

De nombreux acteurs ont donc un vif intérêt à voir se développer la possibilité d'analyser ces flux pour détecter les informations qui les concernent. Les champs d'application sont nombreux, de l'analyse de tendances marketing à la détection de menaces terroristes, en passant par la recherche de signaux d'alerte pour la pharmacovigilance.

Les technologies de la fouille de textes (ingénierie linguistique, apprentissage automatique) sont donc de plus en plus souvent sollicitées pour analyser les corpus de médias sociaux.

Des problèmes spécifiques pour l'analyse linguistique

Les méthodes développées pendant cinq décennies par l'informatique linguistique pour le traitement automatique des langues ont été soumises à de nouvelles contraintes lorsqu'il s'est agi de les appliquer aux médias sociaux. De nombreuses tâches (reconnaissance d'entités, résolution de chaînes d'anaphores...) impliquent des phases de prétraitement qui fonctionnent de manière optimale avec des corpus de textes constitués de phrases complètes, écrits dans une langue homogène formelle, et avec une orthographe normalisée. Ces conditions ne sont en général pas présentes dans les contenus engendrés par les utilisateurs des médias sociaux. L'exactitude d'une tâche comme l'étiquetage en parties du discours (pour prendre l'exemple d'une tâche éprouvée) tombe de 97 % à 80 % lorsque l'on passe d'un corpus d'articles de journaux à un corpus de *tweets*, microgenre textuel (*microblogging*) limité à cent quarante signes (à l'époque où l'étude a été réalisée), et où pullulent les ellipses et les abréviations.

Pourtant les utilisateurs (habitués) comprennent les *tweets*. Ce qui leur permet de les comprendre est que l'information qui n'est pas présente dans le texte lui-même est présente dans son entour hypertextuel et dans ses composantes multimodales. Un *tweet*, par exemple, est souvent inséré dans une conversation, et le contexte de la conversation n'est pas repris, car il est supposé connu du lecteur. Il est émis par un interlocuteur qui occupe une certaine position dans un graphe, par rapport au lecteur autant que par rapport à d'autres interlocuteurs – que l'on cite, que l'on mentionne, auxquels on répond. Il contient des « mots-dièses » (*hashtags*) qui sont autant d'ancres hypertextuelles dont le but est de se positionner dans un espace de discussion dynamique. Il contient des « émoticones », pictogrammes numériques qui permettent en un seul caractère *Unicode* d'exprimer des sentiments ou des prises de position. Il est accompagné d'images qui véhiculent une partie de l'information, qui certes ne prend sens qu'avec le texte, mais dispense le texte de la développer.

Dans leur chapitre 2, Farzindar et Inkpen ont passé en revue différentes tâches élémentaires qui constituent les « briques de base » de l'ingénierie linguistique (segmentation, étiquetage, *chunking*, analyse syntaxique, détection d'entités nommées, identification de langue) et ont caractérisé, pour chacune d'elles, les adaptations qu'elles doivent subir pour être adaptées aux corpus de textes de médias sociaux.

Ces adaptations peuvent consister en prétraitement des textes eux-mêmes (normalisation de la ponctuation), en annotation de corpus d'entraînement (pour réentraîner les étiqueteurs ou les analyseurs à la « syntaxe Twitter »), en redéfinition des catégories de sortie (adaptation de l'ensemble des catégories de parties du discours pour prendre en compte des éléments non linguistiques tels que mots-dièses, mentions nommées, émoticones, images, URL), ou en réentraînement des paramètres des algorithmes d'apprentissage (par exemple pour l'identification de langue).

Après ce tour d'horizon des phases d'adaptation des techniques existantes pour les textes des réseaux sociaux, le chapitre 3 aborde la question de leur interprétation. Il s'agit, bien sûr, dans ce contexte d'interprétation automatique, du processus consistant à inférer à partir des textes, des représentations formelles qui pourront être utilisées pour l'analyse automatique agrégée de grandes quantités de ces textes. Dans le chapitre 3, les auteurs exposent donc les informations que l'on peut en extraire. Elles montrent, par exemple, que des algorithmes d'apprentissage neuronal profond peuvent localiser, avec un certain degré d'exactitude, les utilisateurs à partir de ce qu'ils écrivent, même lorsque ceux-ci n'autorisent pas leur appareil à partager leurs coordonnées géographiques exactes. Elles expliquent ensuite dans quelle mesure certaines des tâches les plus courantes en analyse d'information peuvent être menées sur des textes de médias sociaux.

L'annotation sémantique consiste à attribuer aux entités détectées dans les textes des étiquettes correspondant aux éléments connus d'une source de connaissances « contrôlée » (thesaurus ou ontologie) : on parle d'*entity linking*. Pour atteindre une certaine efficacité sur des textes de type *microblog*, il faut faire usage de tous les indices possibles (contexte conversationnel, mots-dièses, mentions nommées...) ; encore que le score maximal atteint dans l'état de l'art, celui du système YODIE développé par l'équipe GATE, plafonne-t-il à 0,45 (en termes de F-mesure).

D'autres tâches, comme la détection d'opinion, d'émotion, ou de sarcasme, nécessitent également des adaptations du même type, c'est-à-dire, en résumé, l'utilisation de techniques éprouvées de TAL et d'apprentissage, mais en élargissant le spectre des variables d'entrée pour prendre en compte les liens extérieurs et les relations connues dans le graphe des utilisateurs (par exemple, le sens d'un « bravo ! » est à interpréter différemment s'il est adressé en réponse au message d'un utilisateur contre lequel il existe un historique d'opposition polémique). L'état de l'art sur les tâches de détection d'événements ou de sujets, de résumé automatique, et de traduction automatique, est également exploré.

Une large gamme d'applications

Dans leur chapitre 4, les auteurs passent en revue les différentes applications possibles de l'analyse de corpus de médias sociaux. Celles auxquelles est consacrée une section sont celles qui font l'objet d'un corpus de recherche abondant et dynamique.

Dans le domaine de la santé, l'analyse des réseaux sociaux peut être utile pour la pharmacovigilance (en permettant de détecter les effets secondaires dont se plaignent spontanément les utilisateurs) ; une nouvelle partie a également été ajoutée dans cette deuxième édition de l'ouvrage pour présenter l'application de détection de signaux d'alerte de la dépression ou d'envies de suicide.

Dans le domaine de l'analyse financière, ce que disent les utilisateurs sur Twitter peut être un indicateur du moral des consommateurs, ou de celui des investisseurs. Les opinions exprimées sur des entreprises sont également corrélées à leur valeur d'investissement, qu'elles en soient le reflet ou – en partie – la cause : une étude a

montré que des jugements (positifs ou négatifs) exprimés sur Twitter se reflétaient dans un délai d'un à dix jours sur la valorisation boursière des entreprises.

Les textes des réseaux sociaux peuvent également être utilisés pour prédire les intentions de vote, faire de la mercatique en ligne, ou plus généralement modéliser certaines caractéristiques de la personnalité de l'utilisateur pour toutes sortes de buts. Enfin, la surveillance des sujets émergents de façon massive et plus ou moins soudaine dans une zone précisément localisée peut donner des indices forts sur la survenue d'événements catastrophiques : raz-de-marée, tremblement de terre, pollution, ou attaque terroriste. Ces informations peuvent être corrélées à d'autres sources de données (sismographes, capteurs de qualité de l'air...), mais n'en restent pas moins utiles : la pénétration des *smartphones* dans l'Inde du Nord est bien plus élevée que celle des capteurs de qualité de l'air. Quant à la surveillance du terrorisme, l'analyse des productions des utilisateurs sur les réseaux sociaux ne permet pas seulement de savoir quand une attaque survient, elle peut également servir à détecter des signes précurseurs de la radicalisation d'un utilisateur (focalisation sur certains types de sujets, changement dans la typologie des utilisateurs fréquentés, adoption d'un vocabulaire marqué).

Enfin, le chapitre 5 examine les problèmes spécifiques que posent le recueil de corpus de textes dans les médias sociaux (problèmes techniques de débit et de volume, mais aussi problèmes de vie privée et de confidentialité des données), puis l'annotation, et enfin l'évaluation, avec un tour d'horizon des différents *benchmarks* d'évaluation qui ont été conçus spécialement pour les systèmes d'analyse des réseaux sociaux.

Un état de l'art utile en 2018

L'ouvrage de Farzindar et Inkpen offre, en somme, un état de l'art particulièrement utile et bien informé de l'avancée des travaux sur le domaine de l'analyse des médias sociaux (tout particulièrement centré sur Twitter). De nombreuses références, soigneusement mises à jour pour cette deuxième édition, donnent un aperçu panoramique des travaux actuels.

Le lecteur qui y chercherait une réflexion plus approfondie, en termes d'analyse linguistique des nouveaux genres textuels créés par ces pratiques d'échanges, risquerait de rester sur sa faim. Il y aurait beaucoup de choses à dire sur les modes d'interprétation de ces genres émergents, comportant moins de contenu textuel *in praesentia* et beaucoup plus de références hypertextuelles et intermodales. Le livre *Natural Language Processing for Social Media* se cantonne aux technologies du langage telles qu'elles étaient fin 2017. Il a sur ce plan, en tant qu'état de l'art à une époque bien précise, une utilité incontestable.

Résumés de thèses

Rubrique préparée par Sylvain Pogodalla

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
sylvain.pogodalla@inria.fr

Elizaveta CHERNYSHOVA : elizaveta.chernyshova@gmail.com

Titre : Expliciter et inférer dans la conversation. Modélisation de la séquence d'explicitation dans l'interaction

Mots-clés : Implicite, explicite, inférence, interaction, analyse conversationnelle, modélisation.

Title: *Making Explicit and Inferring in Conversations. A Model of Explicitation Sequence in Interaction*

Keywords: *Implicit, explicit, inference, interaction, conversation analysis, modeling.*

Thèse de doctorat en sciences du langage, UFR Sciences du Langage, ICAR, UMR 5191, Université Lumière Lyon 2. Thèse soutenue le 17/12/2018.

Jury : Mme Véronique Traverso (DR, CNRS, codirectrice), M. Sylvain Kahane (Pr, Université Paris Ouest Nanterre la Défense, codirecteur), Mme Claire Beyssade (Pr, Université Paris 8, rapporteur), Mme Maj-Britt Mosegaard Hansen (Pr, University of Manchester, Royaume-Uni, rapporteur), Mme Nathalie Rossi-Gensane (Pr, Université Lumière Lyon 2, présidente), M. Arnulf Deppermann (Pr, Universität Mannheim, Allemagne, examinateur).

Résumé : *Cette thèse porte sur la co-construction de la signification en interaction et les manifestations des processus interprétatifs des participants. En s'intéressant au processus d'explicitation, c'est-à-dire le processus par lequel un contenu informationnel devient explicite dans la conversation, elle propose une étude pluridimensionnelle de la séquence conversationnelle en jeu dans ce processus. La co-construction de la signification est ici abordée comme relevant d'une transformation informationnelle et de l'inférence.*

Nos analyses ont porté sur un corpus de français parlé en interaction, en contexte de repas et apéritifs entre amis. À partir d'une collection de séquences d'explicitation, définies comme des configurations dans lesquelles une inférence est soumise à validation, ce travail propose une analyse multidimensionnelle, portant un double regard sur les données : celui de l'analyse conversationnelle et celui d'une modélisation de la pratique d'explicitation. Ainsi, nous proposons de parcourir cette pratique selon trois axes d'analyse : (a) une analyse séquentielle, s'intéressant au déploiement de la séquence d'explicitation et des éléments la composant ; (b) une analyse reposant sur une modélisation de la gestion informationnelle dans cette séquence ; et (c) une analyse des formats linguistiques employés pour l'exhibition du processus inférentiel. Un des enjeux de ce travail est l'élaboration d'un modèle conversationnaliste pour la gestion informationnelle et son application à l'analyse des données de langue parlée en interaction.

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-02070720

Arnaud FERRÉ : arnaud.ferre.pro@gmail.com

Titre : Représentations vectorielles et apprentissage automatique pour l'alignement d'entités textuelles et de concepts d'ontologie : application à la biologie

Mots-clés : Extraction d'information, normalisation, plongement lexical, intelligence artificielle, traitement automatique des langues.

Title: *Vector Representations and Machine Learning for Alignment of Text Entities with Ontology Concepts: Application to Biology*

Keywords: *Information extraction, normalization, word embedding, artificial intelligence, natural language processing.*

Thèse de doctorat en informatique, Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI), UPR 3251, département STIC, Université Paris-Sud, Orsay, sous la direction de Claire Nédellec (DR, INRA), Pierre Zweigenbaum (DR, CNRS, LIMSI, Orsay). Thèse soutenue le 24/05/2019.

Jury : Mme Claire Nédellec (DR, INRA, codirectrice), M. Pierre Zweigenbaum (DR, CNRS, LIMSI, Orsay, codirecteur), M. Alexandre Allauzen (Pr, Université Paris-Sud, LIMSI, Orsay, président), Mme Nathalie Aussenac (DR, CNRS, IRIT, Toulouse, rapporteur), M. Emmanuel Morin (Pr, Université de Nantes, LS2N, rapporteur), M. Vincent Claveau (CR, CNRS, IRISA, examinateur).

Résumé : *L'augmentation considérable de la quantité des données textuelles rend aujourd'hui difficile leur analyse sans l'assistance d'outils. Or, un texte rédigé en langue naturelle est une donnée non structurée, c'est-à-dire qu'elle n'est interprétable que par un programme informatique spécialisé, sans lequel les informations des textes*

restent largement sous-exploitées. Parmi les outils d'extraction automatique d'information, nous nous intéressons aux méthodes d'interprétation automatique de textes pour la tâche de normalisation d'entité, qui consiste en la mise en correspondance automatique des mentions d'entités de textes avec des concepts d'un référentiel.

Pour réaliser cette tâche, nous proposons une nouvelle approche par alignement de deux types de représentations vectorielles d'entités capturant une partie de leur sens : les plongements lexicaux pour les mentions textuelles et des « plongements ontologiques » pour les concepts, conçus spécifiquement pour ce travail. L'alignement entre les deux se fait par apprentissage supervisé. Les méthodes développées ont été évaluées avec un jeu de données de référence du domaine biologique et elles représentent aujourd'hui l'état de l'art pour ce jeu de données. Ces méthodes sont intégrées dans une suite logicielle de traitement automatique des langues et les codes sont partagés librement.

URL où le mémoire peut être téléchargé :
https://tel.archives-ouvertes.fr/tel-02166253

Natalia GRABAR : natalia.grabar@univ-lille.fr

Titre : Adaptation de documents techniques pour les locuteurs non spécialisés

Mots-clés : Simplification, domaine médical, acquisition de ressources, apprentissage automatique.

Title: *Adaptation of Technical Documents for Non-Specialized Speakers*

Keywords: *Simplification, medical domain, acquisition of resources, machine learning.*

Habilitation à diriger des recherches en informatique, Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI), UPR 3251, Université Paris-Sud, Orsay, sous la direction de François Yvon (DR, CNRS). Habilitation soutenue le 17/05/2019.

Jury : M. François Yvon (DR, CNRS, directeur), Mme Pascale Sébillot (Pr, INSA de Rennes, rapporteur), M. Cédrick Fairon (Pr, Université Catholique de Louvain, Belgique, rapporteur), Mme Pierrette Bouillon (Pr, Université de Genève, Suisse, rapporteur), Mme Chantal Reynaud (Pr, Université Paris-Sud, présidente), M. Stefan Schulz (Pr, Graz General Hospital and University Clinics, Autriche, examinateur), M. Nabil Hathout (DR, CNRS, examinateur).

Résumé : *Comme tout domaine de spécialité, le domaine médical manipule des notions très spécifiques (blépharospasme, alexitymie, appendicectomie), qui sont difficiles à comprendre par les non-spécialistes. Nous proposons un ensemble de travaux dont l'objectif général consiste à adapter les documents techniques de santé et à en assurer une meilleure compréhension par les non-spécialistes. Pour atteindre cet ob-*

jectif, nous proposons une série d'expériences qui font partie d'un processus complexe et ambitieux : (1) la catégorisation des documents selon la difficulté qu'ils présentent ; (2) la détection de passages difficiles au sein des documents ; (3) l'acquisition de ressources pour la simplification lexicale et sémantique des documents ; (4) l'alignement de phrases parallèles à partir de corpus comparables pour engendrer des règles de transformation syntaxique. De plus, une partie non expérimentale du travail est dédiée à l'analyse des travaux de l'état de l'art autour de l'évaluation de la simplification de documents. De manière générale, la recherche que nous présentons ici est une recherche appliquée, motivée par des besoins réels. Chaque étape est effectuée avec une méthode clairement décrite et testée, dont les résultats sont évalués, positionnés par rapport à l'état de l'art et discutés. En fonction des étapes et des tâches, différentes méthodes sont exploitées (à base de règles, par apprentissage supervisé, avec ou sans connaissances linguistiques...). À différentes étapes de ce travail, il a également été nécessaire de construire de nouvelles ressources (lexique, corpus...) dont la genèse est également retracée. En dehors de la simplification lexicale et de la compréhension de textes de spécialité, les résultats et ressources obtenus peuvent être utiles pour d'autres applications et tâches du traitement automatique des langues (TAL) : recherche et extraction d'information, systèmes de questions-réponses, implication textuelle...

URL où le mémoire peut être téléchargé :
http://natalia.grabar.free.fr/publications/grabar-HDR2019.pdf

---

Aurélie NÉVÉOL : aurelie.neveol@limsi.fr

Titre : Traitement automatique de la langue biomédicale

Mots-clés : Extraction d'information, représentation des connaissances, recherche translationnelle.

Title: *Biomedical Natural Language Processing*

Keywords: *Information extraction, knowledge representation, translational research.*

Habilitation à diriger des recherches en informatique, UFR d'informatique, Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI), UPR 3251, CNRS, Université Paris-Sud, Orsay, sous la direction de Pierre Zweigenbaum (DR, CNRS, LIMSI, Orsay). Habilitation soutenue le 26/11/2018.

Jury : M. Pierre Zweigenbaum (DR, CNRS, LIMSI, Orsay, directeur), M. Marc Cuggia (Pr et praticien hospitalier, CHU Pontchaillou, Université de Rennes 1, examinateur), M. Cédrick Fairon (Pr, Université Catholique de Louvain, Belgique, rapporteur), Mme Christine Froidevaux (Pr, Université Paris-Sud, présidente), M. Emmanuel Morin (Pr, Université de Nantes, LS2N, rapporteur), Mme Lynda Tamine Lechani (Pr, Université Paul Sabatier, Toulouse, rapporteur).

Résumé : *Dans le domaine biomédical, les informations cliniques et institutionnelles sont contenues dans le texte de publications scientifiques ou de dossiers patients et ne sont pas directement accessibles à des fins de traitement automatique. Pour pallier cela, le traitement automatique de la langue naturelle peut offrir des méthodes d'extraction d'information afin de convertir des textes libres en représentations exploitables pour la recherche médicale et la santé publique.*

Cependant, ces méthodes doivent être robustes face au volume, à la technicité et à la diversité des textes à traiter.

Le traitement automatique de la langue biomédicale (ou TAL biomédical) est un champ de recherche pluridisciplinaire qui mobilise l'informatique, la linguistique ainsi que la médecine. Il s'inscrit dans le champ du traitement automatique de la langue, tout en allant au-delà du service rendu à la médecine.

Trois thématiques ont particulièrement fait l'objet de mon travail ces dernières années : (1) la modélisation des informations ; (2) l'analyse de textes en langue de spécialité ; et (3) les applications biomédicales concrètes.

Tout ce travail repose sur l'analyse de corpus variés du domaine. Ainsi, le développement de ressources en soutien du TAL biomédical, en particulier pour les langues autres que l'anglais comparativement peu dotées, est un défi scientifique majeur. Mes contributions sur ce point s'appuient sur une analyse des schémas de représentation des connaissances dans le domaine, qui a permis le développement de corpus annotés destinés à être partagés par la communauté à des fins de développement méthodologique et d'évaluation. Une autre série de contributions a porté sur la proposition de méthodes d'analyse de textes médicaux, par exemple avec l'extraction d'entités et de relations. Ce travail a permis de montrer l'importance de la question de l'adaptation en domaine et de l'adaptation multilingue. Enfin, mes contributions à des études ciblées sur des applications biomédicales permettent de souligner l'impact attendu du traitement automatique de la langue en épidémiologie, en santé publique et sur les pratiques de recherche au-delà de ces disciplines. L'un des défis du TAL biomédical est de réaliser pleinement ce potentiel en mettant des outils à disposition de la communauté médicale afin de devenir un levier incontournable de la recherche translationnelle.

Ces travaux ont été réalisés en collaboration avec de nombreux collègues notamment au LIMSI CNRS et à la U.S. National Library of Medicine dans le cadre de plusieurs thèses, post-docs, stages de masters, et pour certains dans le cadre de projets de recherche ANR, H2020 et Digicosme, impliquant différents organismes, tels que l'Inserm, le CEA et des partenaires hospitaliers.

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-02167096