Traitement automatique des langues

Traitement automatique des langues peu dotées

sous la direction de
Delphine Bernhard
Claudia Soria

Vol. 59 - n°3 / 2018

ATALA

Traitement automatique des langues peu dotées

ATALA

Revue de
l'Association
pour le Traitement
Automatique
des Langues

TAL
Vol.
59

n°3
2018

Traitement automatique des langues peu dotées

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

©ATALA, 2018

ISSN 1965-0906

https ://www.atala.org/revuetal

Le Code de la propriété intellectuelle n'autorisant, aux termes de l'article L. 122-5, d'une part, que les « copies ou reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective » et, d'autre part, que les analyses et les courtes citations dans un but d'exemple et d'illustration, « toute représentation ou reproduction intégrale, ou partielle, faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause, est illicite » (article L. 122-4).
Cette représentation ou reproduction, par quelque procédé que ce soit, constituerait donc une contrefaçon sanctionnée par les articles L. 225-2 et suivants du Code de la propriété intellectuelle.

Traitement automatique des langues

Comité de rédaction

Rédacteurs en chef

Cécile Fabre - CLLE, Université Toulouse 2
Emmanuel Morin - LS2N, Université Nantes
Sophie Rosset - LIMSI, CNRS
Pascale Sébillot - IRISA, INSA Rennes

Membres

Salah Aït-Mokhtar - Naver Labs Europe, Grenoble
Maxime Amblard - LORIA, Université Lorraine
Frédéric Béchet - LIF, Université Aix-Marseille
Patrice Bellot - LSIS, Université Aix-Marseille
Laurent Besacier - LIG, Université de Grenoble
Pierrette Bouillon - ETI/TIM/ISSCO, Université de Genève, Suisse
Thierry Charnois - LIPN, Université Paris 13
Vincent Claveau - IRISA, CNRS
Mathieu Constant - ATILF, Université Lorraine
Laurence Danlos - ALPAGE, Université Paris 7
Gaël Harry Dias - GREYC, Université Caen Basse-Normandie
Iris Eshkol - MoDyCo, Université Paris Nanterre
Dominique Estival - The MARCS Institute, University of Western Sydney, Australie
Cyril Goutte - Technologies Langagières Interactives, CNRC, Canada
Nabil Hathout - CLLE-ERSS, CNRS
Sylvain Kahane - MoDyCo, Université Paris Nanterre
Mathieu Lafourcade - LIRMM, Université Montpellier 2
Philippe Langlais - RALI, Université de Montréal, Canada
Yves Lepage - Université Waseda, Japon
Denis Maurel - Laboratoire d'Informatique, Université François-Rabelais, Tours
Sien Moens - KU Leuven, Belgique
Philippe Muller - IRIT, Université Paul Sabatier, Toulouse
Alexis Nasr - LIF, Université Aix-Marseille
Adeline Nazarenko - LIPN, Université Paris 13
Patrick Paroubek - LIMSI, CNRS
Sylvain Pogodalla - LORIA, INRIA
François Yvon - LIMSI, Université Paris Sud

Secrétaire

Marco Dinarelli - LIG, CNRS

Traitement automatique des langues

Volume 59 – n°3 / 2018

TRAITEMENT AUTOMATIQUE DES LANGUES PEU DOTÉES

Table des matières

Traitement automatique des langues peu dotées

Delphine Bernhard* — **Claudia Soria****

* *LiLPa, Université de Strasbourg*
** *CNR-ILC, Pisa*

RÉSUMÉ. *Jusqu'à récemment, la plupart des travaux de recherche sur le traitement automatique des langues (TAL) étaient axés sur quelques langues bien décrites avec de nombreux locuteurs. La situation évolue rapidement, avec une nette augmentation de l'intérêt pour les langues dites « sous-dotées ». L'objectif de ce numéro de la revue* Traitement automatique des langues *est de donner un aperçu des recherches en cours sur le TAL pour des langues peu dotées du monde entier, couvrant une grande variété de tâches. Les articles sélectionnés portent à la fois sur des langues qui sont encore au début du processus et sur des langues dont la situation s'est très récemment améliorée. Nous espérons qu'ils pourront servir à orienter les recherches futures sur d'autres langues disposant de peu de ressources et d'outils.*

ABSTRACT. *Until recently, most of the research work in Natural Language Processing (NLP) has been focused on a few well-described languages with many speakers. The situation is rapidly evolving, with a clear increase in the interest towards so called "under-resourced" languages. The goal of this issue of the Traitement Automatique des Langues journal is to give an overview of current research on NLP for under-resourced languages from all over the world, encompassing a large variety of tasks. The selected papers address languages which are still at very early stages as well as languages whose situation has very recently improved. We hope that they can be helpful to guide future research on other languages with little or no resources and tools.*

MOTS-CLÉS : *langues peu dotées, ressources linguistiques, outils de TAL.*

KEYWORDS: *under-resourced languages, language resources, NLP tools.*

TAL. Volume 59 – n°3/2018, pages 7 à 14

1. Introduction

Jusqu'à récemment, la plupart des travaux de recherche en traitement automatique des langues (TAL) se sont concentrés sur quelques langues bien décrites et ayant de nombreux locuteurs (Del Gratta *et al.*, 2014). Le manque d'intérêt pour d'autres langues et variétés linguistiques «sous-dotées» peut s'expliquer par différentes raisons : manque de financement, de ressources humaines, de technologie appropriée, de descriptions linguistiques complètes et précises, de reconnaissance académique par la communauté scientifique, pour n'en nommer que quelques-unes. Les langues sous-dotées posent néanmoins d'importants défis scientifiques qui ouvrent des pistes de progrès pour le TAL en général.

Premièrement, à une époque où les méthodes de l'état de l'art nécessitent généralement de grandes quantités de données annotées, le travail sur des langues sous-dotées impose souvent des méthodes capables de traiter des jeux de données de petite taille (*small data*). Par exemple, les plus petits corpus arborés issus des Universal Dependencies ne contiennent que quelques milliers de *tokens* : environ 1 000 pour l'akkadien ou le sanskrit, 10 000 pour le breton ou le féroïen (pour ne citer que quelques langues), contre plus de 1 million pour l'arabe ou le français (Universal Dependencies, 2018). Des méthodes efficaces et fiables pour l'acquisition et la collecte de ressources et d'annotations (OCR, *crowdsourcing*, médias sociaux, etc.) sont essentielles (McShane *et al.*, 2002). De plus, les outils d'annotation automatique doivent être capables de gérer le manque de données (Abraham *et al.*, 2016), les problèmes de qualité et les mots hors vocabulaire (Liu et Kirchhoff, 2018). Ces dernières années, des modèles s'appuyant sur la projection d'annotations à partir d'autres langues ont vu le jour (Sukhareva et Chiarcos, 2014 ; Agić *et al.*, 2016). Ces méthodes tirent parti des outils d'annotation existants ou des ressources annotées pour des langues mieux dotées, grâce à l'utilisation de corpus parallèles. Elles dépendent donc largement de la qualité de l'alignement que l'on peut obtenir (Akbik et Vollgraf, 2018).

Deuxièmement, compte tenu des difficultés à trouver des ressources telles que des lexiques ou des corpus, les données collectées sont souvent très hétérogènes et correspondent à différentes époques, aires linguistiques ou domaines, par exemple des corpus de textes intégrant différentes variétés géolinguistiques et portant sur différents sujets à différentes époques (Goldhahn *et al.*, 2016). Cette hétérogénéité implique aussi souvent des variations dans la graphie, dues soit à une évolution des normes orthographiques dans le temps, soit à l'absence de normes orthographiques pour les langues ou les variétés linguistiques qui sont essentiellement orales et rarement écrites (Kurimo *et al.*, 2017). Dans de tels cas, la normalisation orthographique est souvent la solution préférée (Samardzic *et al.*, 2015). En outre, l'alternance codique peut également causer des problèmes, ce qui nécessite d'identifier la langue ou la variété de langue (Lavergne *et al.*, 2014).

Troisièmement, les travaux de TAL pour les langues sous-dotées ont tendance à être réalisés dans des groupes de recherche isolés ou dispersés, et les ressources produites utilisent souvent des formats et des normes différents. Trouver ces ressources, y

accéder et les rendre interopérables pour qu'elles puissent être réutilisées peut devenir un défi en soi. Quand il s'agit de langues sous-dotées, les questions d'interopérabilité des données et des métadonnées deviennent d'une importance cruciale pour combiner et réutiliser les quelques ressources et outils qui pourraient être disponibles (Alegria *et al.*, 2011).

L'objectif de ce numéro de *Traitement automatique des langues* est de donner un aperçu de la recherche actuelle sur le TAL pour les langues peu dotées du monde entier, englobant une grande variété de tâches. Nous avons reçu 23 soumissions en tout, ce qui montre que le sujet abordé dans ce numéro spécial est particulièrement pertinent et figure en bonne place parmi les thématiques de recherche actuelles. Il ne s'agit plus d'un sujet de recherche très spécialisé et secondaire. C'est ce qu'attestent également les nombreuses conférences et ateliers consacrés au TAL pour les langues peu dotées ces dernières années : par exemple la série d'ateliers CCURL (*Collaboration and Computing for Under-Resourced Languages*) à partir de 2014, SLTU (*Spoken Language Technologies for Under-Resourced Languages*) à partir de 2008 ou le « *Less-Resourced Languages Workshop* » organisé depuis 2009 à la conférence L&TC. Les principales conférences du domaine, telles que COLING ou ACL, proposent régulièrement des sessions dédiées aux langues moins dotées. En 2017, un groupe d'intérêt spécial ELRA-ISCA sur les langues sous-dotées (SIGUL) a été créé. Les défis posés par les langues peu dotées sont de plus en plus souvent abordés dans les communications présentées lors de conférences et d'ateliers généralistes, en particulier pour évaluer les méthodes face au manque de données, par exemple « *CoNLL 2018 UD Shared Task* » (Zeman *et al.*, 2018), la campagne d'évaluation IWSLT 2018 (Jan *et al.*, 2018), ou l'atelier « *Workshop on Deep Learning Approaches for Low-Resource NLP* » à ACL 2018 (Haffari *et al.*, 2018). Les langues peu dotées fournissent en effet des données d'évaluation difficiles, mais réalistes, pour les méthodes état de l'art.

Plus important encore, le fait que des travaux de recherche universitaire de grande qualité soient menés pour les langues peu dotées pourrait être lié à une prise de conscience partagée et de plus en plus répandue de l'importance de la représentation numérique pour toutes les langues. Alors que la numérisation de la société moderne s'accentue et que la fracture numérique entre les langues très utilisées et les langues moins répandues se creuse, les locuteurs de langues minoritaires ou moins répandues sont confrontés à des disparités dans l'accès à l'information. La disponibilité de contenus dans de nombreuses langues différentes réduirait certainement cette inégalité.

Le développement des technologies TAL et des ressources linguistiques pour les langues qui ont été ou sont encore exclues du monde numérique est une condition préalable pour qu'elles soient pleinement utilisables sur les médias numériques dans un avenir que l'on n'espère pas si lointain. Cela garantira une meilleure égalité des droits numériques à tous les citoyens, qui pourraient ainsi bénéficier d'un environnement favorable pour pouvoir s'exprimer et créer leur propre contenu culturel dans les langues locales, un pas de plus vers une diversité linguistique et culturelle plus large et plus forte.

2. Résumés des articles

Le présent numéro se compose de quatre articles traitant de langues à différents stades de développement concernant les ressources et les outils pour le TAL.

Les deux premiers articles se concentrent sur l'acquisition de données annotées, en s'appuyant sur deux approches différentes : une campagne d'annotation « classique » réalisée par des annotateurs formés pour l'annotation du serbe, et une approche reposant sur le *crowdsourcing* pour deux langues de France, l'alsacien et le créole guadeloupéen.

Le troisième article porte sur l'analyse syntaxique de deux langues à faibles ressources, le same du nord et le komi-zyriène.

Enfin, le dernier article donne un aperçu de l'état de l'art du traitement automatique du dialecte tunisien, qui recense les ressources et outils disponibles.

2.1. *De la constitution d'un corpus arboré à l'analyse syntaxique du serbe*

L'article décrit le processus de production d'un corpus d'environ 101 000 *tokens* en serbe, annoté avec des propriétés morphosyntaxiques, les lemmes et les relations de dépendance syntaxique. Avant les travaux décrits dans l'article, le serbe a été doté de certaines ressources et outils, mais la plupart d'entre eux ne sont pas librement et facilement disponibles (à l'exception notable de certains travaux récents : un lexique morphosyntaxique, publié en 2016, et le corpus Universal Dependencies qui a été publié en 2017). La méthodologie proposée dans cet article vise à optimiser les ressources linguistiques et humaines limitées en adaptant les ressources existantes d'une langue très proche (le croate), en utilisant des ressources lexicales qui ont été produites de manière collaborative (Wiktionary) et en préannotant automatiquement le corpus.

2.2. *À l'écoute des locuteurs : production participative de ressources langagières pour des langues non standardisées*

L'utilisation du *crowdsourcing* pour l'annotation est encore relativement inexplorée pour les langues sous-dotées. Dans cet article, une méthodologie d'annotation en parties du discours s'appuyant sur le *crowdsourcing* est présentée et appliquée aux dialectes alsaciens et au créole guadeloupéen. Dans les deux cas, les ressources existantes sont très limitées et la graphie n'est pas normalisée. Deux plateformes participatives différentes sont présentées. La première vise à obtenir des annotations en parties du discours. Les corpus sont préannotés avant d'être présentés aux participants. Les corpus qui en résultent sont ensuite utilisés pour entraîner des étiqueteurs. La deuxième plateforme vise la collecte de corpus bruts (en se concentrant pour l'instant sur les recettes de cuisine) ainsi que des lexiques de variantes graphiques et dialectales. Les utilisateurs peuvent également corriger des annotations morphosyntaxiques qui ont été produites automatiquement.

2.3. *Analyse syntaxique de langues faiblement dotées à partir de plongements de mots multilingues*

Le same du nord et le komi-zyriène sont deux langues finno-ougriennes sous-dotées qui n'ont pas les mêmes niveaux de développement en ce qui concerne les ressources et les outils de TAL : alors que le same du nord dispose de lexiques flexionnels relativement complets, ainsi que d'un corpus Universal Dependencies, le komi-zyriène manque de telles ressources. L'approche proposée pour l'analyse automatique de ces langues est multilingue et ne nécessite qu'un petit lexique bilingue et une annotation syntaxique manuelle de quelques phrases. La méthode utilise également des plongements de mots, à la fois pour les langues cibles et pour des langues mieux dotées en ressources (finnois ou russe), ainsi que des corpus Universal Dependencies existants (anglais, finnois ou russe).

2.4. *Un état de l'art du traitement automatique du dialecte tunisien*

Le dialecte tunisien est l'un des nombreux dialectes parlés dans les pays arabes. Il diffère considérablement de l'arabe standard moderne, qui a été beaucoup plus étudié et doté de nombreuses ressources et outils. L'article passe en revue les travaux récents visant à collecter des ressources et des outils pour le dialecte tunisien. Il se concentre sur les aspects suivants : corpus (transcriptions orales, Web, corpus parallèles), lexiques, ontologies, traitement de la parole, outils d'analyse morpho-syntaxique, identification de la langue, traduction, analyse des sentiments, normalisation.

La plupart des travaux décrits sont très récents, ce qui montre que la communauté des chercheurs s'est récemment beaucoup intéressée au développement de ressources et d'outils pour le dialecte tunisien. Cette implication a permis une nette amélioration de la situation. Cependant, l'article note que seulement 24 % des ressources énumérées sont téléchargeables gratuitement en ligne, et uniquement deux outils. De plus, les ressources sont encore assez petites et souvent limitées à un domaine spécifique. En conséquence, l'effort de construction de ressources et d'outils pour le dialecte tunisien doit se poursuivre.

3. Conclusion

Ce numéro est consacré aux langues peu et sous-dotées. Ces termes n'ont pas encore de définition précise et se recoupent largement avec ceux des langues minoritaires et en danger. Ce qui est clair, c'est que ces termes peuvent s'appliquer à un large éventail de langues, à différents stades d'avancement en ce qui concerne les ressources et les outils du TAL. Les langues sous-dotées peuvent être minoritaires ou menacées, mais l'inverse ne s'applique pas toujours (par exemple, le catalan est une langue minoritaire en Espagne, mais ne dispose pas de moins de ressources). D'un autre côté, les langues comptant des millions de locuteurs, et vitales comme l'ourdou

ou certaines langues chinoises, sont également sous-dotées. Appliquer le terme à une langue implique de savoir si la langue considérée dispose ou non des ressources et de la technologie nécessaires pour accéder aux médias numériques comme les autres.

Deux articles de ce numéro traitent de langues qui n'en sont qu'à leurs débuts (l'alsacien, le créole guadeloupéen, le same du nord, le komi-zyriène) et peuvent être utiles pour orienter les recherches futures sur d'autres langues avec peu ou pas de ressources. Les deux autres articles concernent des langues qui ne peuvent probablement plus être considérées comme des langues peu dotées (serbe, dialecte tunisien), mais qui ont très récemment amélioré leur situation. Tous ces exemples montrent comment le statut d'une langue peut non seulement être amélioré, mais aussi conduire à des solutions innovantes. Nous espérons que ces articles intéresseront non seulement les chercheurs, mais aussi les institutions qui prennent des décisions en matière de financement et de politiques linguistiques, afin d'encourager la recherche sur les nombreuses langues qui manquent encore de ressources.

Remerciements

Nous aimerions remercier les rédacteurs en chef et le comité de rédaction de la revue TAL d'avoir eu l'idée d'un numéro spécial sur le thème des langues peu dotées. Nous remercions également les relectrices et relecteurs pour leurs précieux commentaires.

Membres du comité scientifique (par ordre alphabétique) : Gilles Adda (LIMSI-CNRS, France), Antti Arppe (University of Alberta, Canada), Vincent Berment (INALCO, France), Myriam Bras (Université Toulouse Jean Jaurès / CLLE-ERSS, France), Thierry Declerck (DFKI, Allemagne), Chantal Enguehard (LS2N, Nantes, France), Vera Ferreira (CIDLeS - Interdisciplinary Centre for Social and Language Documentation, Portugal), Karën Fort (Université Paris-Sorbonne, France), András Kornai (Hungarian Academy of Sciences, Hongrie), Anne-Laure Ligozat (ENSIIE / LIMSI-CNRS, France), Teresa Lynn (ADAPT DCU, Irlande), Mathieu Mangeot-Nagata (Université de Savoie / LIG, France), Joseph Mariani (LIMSI-CNRS, France), Damien Nouvel (INALCO), Thierry Poibeau (LATTICE-CNRS, France), Laurette Pretorius (University of South Africa, Afrique du Sud), Benoît Sagot (Inria Paris, France), Sakriani Sakti (NAIST, Japon), Kevin Scannell (Saint Louis University, Missouri, Etats-Unis), Yves Scherrer (University of Helsinki, Finlande), Jörg Tiedemann (University of Helsinki, Finlande), Trond Trosterud (Tromsø University, Norvège), Francis Tyers (Moscow Higher School of Economics, Russie), Assaf Urieli (Université Toulouse Jean Jaurès / CLLE-ERSS et Joliciel Informatique, France).

4. Bibliographie

Abraham B., Umesh S., Joy N., « Overcoming Data Sparsity in Acoustic Modeling of Low-Resource Language by Borrowing Data and Model Parameters from High-Resource Languages », *Proceedings of the 17th Annual Conference of the International Speech Communi-*

cation Association (INTERSPEECH 2016) : Understanding Speech Processing in Humans and Machines, p. 3037-3041, 2016.

Agić Ž., Johannsen A., Plank B., Alonso H. M., Schluter N., Søgaard A., « Multilingual Projection for Parsing Truly Low-Resource Languages », *Transactions of the Association for Computational Linguistics*, vol. 4, p. 301-312, 2016.

Akbik A., Vollgraf R., « ZAP : An Open-Source Multilingual Annotation Projection Framework », *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Alegria I., Artola X., de Ilarraza A. D., Sarasola K., « Strategies to develop Language Technologies for Less-Resourced Languages based on the case of Basque », *HAL Id : artxibo-00783393*, 2011.

Del Gratta R., Frontini F., Khan A. F., Mariani J., Soria C., « The LREMap for Under-Resourced Languages », *Proceedings of CCURL 2014 : Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era*, p. 78-83, 2014.

Goldhahn D., Sumalvico M., Quasthoff U., « Corpus collection for under-resourced languages with more than one million speakers », *Proceedings of the LREC Workshop "CCURL 2016 Collaboration and Computing for Under-Resourced Languages : Towards an Alliance for Digital Language Diversity"*, Portorož, Slovenia, 2016.

Haffari R., Cherry C., Foster G., Khadivi S., Salehi B., *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, 2018.

Jan N., Cattoni R., Sebastian S., Cettolo M., Turchi M., Federico M., « The IWSLT 2018 Evaluation Campaign », *Proceedings of the International Workshop on Spoken Language Translation*, p. 2-6, 2018.

Kurimo M., Enarvi S., Tilk O., Varjokallio M., Mansikkaniemi A., Alumaë T., « Modeling under-resourced languages for speech recognition », *Language Resources and Evaluation*, vol. 51, n° 4, p. 961-987, 2017.

Lavergne T., Adda G., Adda-Decker M., Lamel L., « Automatic Language Identity Tagging on Word and Sentence-Level in Multilingual Text Sources : a Case-Study on Luxembourgish », *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014.

Liu A., Kirchhoff K., « Context Models for OOV Word Translation in Low-Resource Languages », *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1 : Research Papers)*, p. 54-67, 2018.

McShane M., Nirenburg S., Cowie J., Zacharski R., « Embedding knowledge elicitation and MT systems within a single architecture », *Machine Translation*, n° 17, p. 271-305, 2002.

Samardzic T., Scherrer Y., Glaser E., « Normalising orthographic and dialectal variants for the automatic processing of Swiss German », *Proceedings of the 7th Language and Technology Conference*, Poznan, 2015.

Sukhareva M., Chiarcos C., « Diachronic proximity vs. data sparsity in cross-lingual parser projection. A case study on Germanic », *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, p. 11-20, 2014.

Universal Dependencies, `http://universaldependencies.org/`, 2018. Accédé le 9 décembre 2018.

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 », *Procee-*

dings of the CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies, p. 1-21, 2018.

De la constitution d'un corpus arboré à l'analyse syntaxique du serbe

Aleksandra Miletic* — **Cécile Fabre*** — **Dejan Stosic***

* *CLLE, Université de Toulouse, CNRS, UT2J, France*
aleksandra.miletic@univ-tlse2.fr

RÉSUMÉ. *Cet article retrace une expérience de constitution d'un corpus arboré pour le serbe, conçu dans le but de doter cette langue des instruments nécessaires à l'analyse syntaxique et, plus généralement, de favoriser des recherches plus systématiques aussi bien en TAL (traitement automatique des langues) qu'en linguistique serbe. Au-delà de la description des résultats de ce projet, nous présentons une méthode de confection d'un corpus arboré qui vise à optimiser les ressources, par définition rares, dont on dispose dans le cas d'une langue peu dotée, qu'il s'agisse de moyens matériels (corpus et outils) ou humains. Nous montrons comment tirer au mieux parti de l'existant pour faciliter le travail des annotateurs humains et accélérer l'enrichissement du corpus, tout en garantissant la validité de l'annotation produite. Cette méthode, basée sur des principes transposables à d'autres langues, a vocation à faciliter la création des corpus arborés pour les langues sous-dotées en général.*

ABSTRACT. *In this paper we describe our work on a treebank for Serbian, which aims to provide this language with tools and resources needed for parsing and, more globally, to encourage research on this language both in NLP (natural language processing) and in theoretical linguistics. Beyond the results of this resource-building project, we also provide a description of a treebank-building method that optimizes the limited resources available for an under-resourced language, both from the technical point of view (tools and corpora) and from that of human resources (annotation process). We show how best to take advantage of what is available in order to facilitate the manual work and accelerate the corpus enrichment process, all the while maintaining a high-quality annotation. Being based on language-independent principles, this method should help forward the creation of treebanks for other under-resourced languages.*

MOTS-CLÉS : *corpus arboré, serbe, méthode d'annotation, optimisation, langues à morphologie flexionnelle riche, langues sous-dotées.*

KEYWORDS: *treebank, Serbian, annotation method, optimization, morphologically rich languages, under-resourced languages.*

1. Introduction

Cet article retrace une expérience de constitution d'un corpus arboré pour le serbe, conçue dans le but de doter cette langue des instruments nécessaires à l'analyse syntaxique et, plus généralement, de favoriser des recherches plus systématiques aussi bien en TAL (traitement automatique des langues) qu'en linguistique serbe. Le serbe n'est pas à proprement parler une langue dépourvue de ressources et d'outils (Krstev *et al.*, 2004 ; Vitas et Krstev, 2004 ; Pavlović-Lažetić *et al.*, 2004), et des efforts récents tendent à rattraper le retard pris (Gesmundo et Samardžić, 2012 ; Jakovljević *et al.*, 2014 ; Samardžić *et al.*, 2017). Néanmoins, une partie importante des ressources existantes n'est pas diffusée ou l'est sous des licences restrictives (section 2.2). Par ailleurs, les seules expériences en analyse syntaxique de cette langue sur un corpus serbe (Jakovljević *et al.*, 2014) ont été basées sur un corpus d'apprentissage minimal (7 000 *tokens*) qui n'a pas été distribué. Le corpus arboré présenté ici, publié alors qu'un corpus serbe issu du projet *Universal Dependencies* (dorénavant UD)[1] vient d'être diffusé (Samardžić *et al.*, 2017), est assorti d'un ensemble de ressources (lexiques et modèles d'analyse syntaxique, de lemmatisation et d'étiquetage morphosyntaxique). Quant aux annotations linguistiques du corpus arboré, outre les indications des fonctions syntaxiques et des catégories grammaticales, le corpus est également doté d'une lemmatisation et d'une annotation en traits morphosyntaxiques fins, tels le cas, le genre, le nombre, etc. Cela a été fait dans le souci d'assurer des conditions optimales pour l'analyse syntaxique : en effet, le serbe est une langue à morphologie flexionnelle riche et à ordre des constituants flexible, et de nombreux travaux montrent que l'analyse syntaxique de ces langues est facilitée par l'utilisation de ces deux types d'annotation (Collins *et al.*, 1999 ; Marton *et al.*, 2013). Le corpus, qui contient 101 000 *tokens*, peut être téléchargé à partir de l'adresse suivante : `https://github.com/aleksandra-miletic/serbian-nlp-resources`.

Au-delà de la description des résultats de ce projet, nous présentons dans cet article une méthode de confection d'un corpus arboré qui vise à optimiser les ressources, par définition rares, dont on dispose dans le cas d'une langue peu dotée, qu'il s'agisse de moyens matériels (corpus et outils) ou humains requis pour le processus d'annotation. En effet, malgré l'expansion continuelle des corpus annotés depuis la fin du XX^{e} siècle, la question des standards et des bonnes pratiques quant à leur constitution reste peu abordée. Hovy et Lavid (2010) proposent une schématisation à huit étapes, articulant plusieurs phases d'annotation et d'évaluation. Pustejovsky et Stubbs (2012) proposent un schéma MATTER en sept points, soit modélisation du phénomène et création des guides, annotation, entraînement (*train*), test, évaluation et révision. À la différence de ces approches, qui soumettent l'adaptation du schéma d'annotation aux résultats des outils du TAL, Fort (2016) préconise la séparation de l'élaboration du corpus et des évaluations dans le cadre du TAL afin de garantir une validité plus générale de la ressource. Par ailleurs, Fort (2012) opère une distinction plus nette entre les périodes principales d'une campagne d'annotation, qui s'organise en travail de prépa-

1. `http://universaldependencies.org/`

ration (identification des participants, constitution du corpus, création du guide d'annotation), précampagne (création d'un corpus de référence minimal, formation des annotateurs), campagne (entraînement des annotateurs, annotation proprement dite, mise à jour du guide) et finalisation (publication du corpus). L'auteur identifie également les différents rôles à l'intérieur d'une campagne (le gestionnaire de campagne, l'expert, les annotateurs, l'évaluateur, etc.) et donne des recommandations en ce qui concerne la qualité de l'annotation. Plus précisément, Fort (2012) met en avant la méthode d'annotation agile, définie par Voormann et Gut (2008) et implémentée par Alex *et al.* (2010). Cette approche préconise une organisation cyclique du travail, présente de manière implicite dans les formalisations de Hovy et Lavid (2010) et Pustejovsky et Stubbs (2012) : le corpus est divisé en échantillons qui sont traités tour à tour ; chaque cycle d'annotation est suivi d'une étape d'évaluation, où l'accord interannotateur est calculé, permettant ainsi de contrôler la qualité de l'annotation, mais aussi de relever les problèmes dans les guides d'annotation et d'y remédier.

Dans le but de faciliter le travail des annotateurs humains et d'accélérer l'enrichissement du corpus, tout en garantissant la validité de l'annotation produite, nous adoptons la méthode de Fort (2012) et y intégrons des éléments de l'annotation agile. Plus concrètement, cette méthode est fondée sur plusieurs principes : l'adaptation de ressources existantes pour une langue proche et mieux dotée ; l'utilisation de ressources lexicales produites de façon collaborative ; la conception d'un processus d'annotation qui articule de façon optimale les phases d'annotation humaine et de préannotation automatique.

La section 2 aborde les principales caractéristiques du serbe et l'état de l'art en TAL de cette langue. Dans la section 3, nous présentons la méthode d'annotation adoptée dans sa globalité, pour détailler ensuite le travail de préparation d'outils permettant l'optimisation de l'annotation (section 4), les campagnes d'annotation manuelle (section 5) et le processus de finalisation du corpus (section 6). Enfin, la section 7 présente nos conclusions en rappelant les jalons d'une méthode optimisée pour le développement de ressources pour les langues peu dotées.

2. Le serbe : aperçu du système

Le serbe est une langue slave méridionale, parlée majoritairement en Serbie et dans les pays de l'ex-Yougoslavie par environ 8,7 millions de locuteurs (Keith, 2006). Il dispose d'un système d'écriture phonétique et a recours de façon quasi équivalente aux deux alphabets cyrillique et latin. Dans ce qui suit, nous proposons un aperçu de ses caractéristiques principales pertinentes pour le TAL et présentons ensuite l'état de l'art du TAL serbe.

2.1. *Principales propriétés linguistiques*

Le serbe exhibe toutes les propriétés phares de la famille slave : il dispose d'un système de déclinaisons relativement complexe, l'ordre des constituants est flexible, il n'y a pas d'articles, le système de l'aspect verbal est particulièrement bien développé, et la réalisation du sujet dans la phrase n'est pas obligatoire (il s'agit d'une langue *pro-drop*). Nous nous concentrerons ici sur deux d'entre elles : la morphologie flexionnelle riche et l'ordre des constituants flexible.

2.1.1. *Morphosyntaxe*

Le serbe dispose d'un système de déclinaison à 7 cas (nominatif, génitif, datif, accusatif, vocatif, instrumental et locatif), des marques du nombre (singulier ou pluriel)[2] et du genre (masculin, féminin ou neutre). Ces trois catégories caractérisent aussi bien les noms que les pronoms et les adjectifs. Par ailleurs, les adjectifs portent des marques du degré de comparaison (positif, comparatif ou superlatif). Par conséquent, on considère typiquement qu'un paradigme nominal contient 14 formes, alors qu'un paradigme adjectival en a 126. Quant à la conjugaison, le paradigme verbal prototypique compte au-delà de 120 formes.

Il existe cependant un degré de syncrétisme important (et en partie systématique), notamment dans les paradigmes de déclinaison. Par exemple, pour les noms et pour les adjectifs, le datif et le locatif sont toujours identiques (cf. la forme *detetu* – le datif ou le locatif singulier du nom *dete* 'enfant') ; pour les adjectifs, les formes du pluriel sont identiques pour les trois genres sauf au nominatif et au vocatif (cf. la forme *lepih* – le génitif pluriel du masculin, du féminin ou du neutre).

2.1.2. *Syntaxe*

Comme c'est souvent le cas, cette morphologie flexionnelle riche est accompagnée d'une flexibilité importante dans l'ordre des constituants, le système casuel prenant en charge l'encodage d'une partie des fonctions syntaxiques. Par exemple, les réalisations prototypiques du sujet, de l'objet direct et de l'objet indirect sont respectivement assurées par le nominatif, l'accusatif et le datif (exemple 1).

(1)

Filip	predstavlja	**An-u**	Alan-u
Filip.NOM.SG	présente	Ana.ACC.SG	Alan.DAT.SG

'*Filip* présente **Ana** à Alan'

L'ordre de ces constituants est très variable : même si l'ordre canonique est SVO, les 5 autres variations sont grammaticales, et l'objet indirect dispose d'un degré de

2. Le serbe exhibe également des traces de l'ancien dual (*paucal*), mais ces formes ne connaissent pas un usage systématique et ne sont typiquement pas considérées comme faisant partie du paradigme nominal canonique (Stanojčić et Popović, 2012).

flexibilité comparable. Une phrase simple comme celle donnée ci-dessus peut donc connaître de nombreuses variations (*Filip* predstavlja Alanu **Anu**, *Filip* **Anu** predstavlja Alanu, *Filip* Alanu predstavlja **Anu**, etc.).

Cette flexibilité au niveau syntaxique va jusqu'à autoriser des structures discontinues (cf. exemple 2). Ici, la seule manière de déterminer la fonction syntaxique de l'adjectif *lepu* 'beau' est de faire appel aux traits morphosyntaxiques qui participent à l'accord : comme le cas, le genre et le nombre de cette forme coïncident avec ceux du nom *knjigu* 'livre', c'est celui-ci qui est son gouverneur plutôt que le nom *Filip*.

(2)

Lep-u	je	Filip	knjig-u	kupio.
beau-ACC.SG.F	AUX	Filip.NOM.SG	livre-ACC.SG.F	acheté

'C'est un beau livre que Filip a acheté.'

Le serbe est donc plus riche en formes fléchies et en traits morphosyntaxiques, et plus variable au niveau syntaxique que les langues telles que l'anglais ou le français. Ces propriétés ont un effet concret sur le traitement automatique, discuté dans la section suivante.

2.2. *Traitement automatique du serbe et des langues proches*

Les langues comme le serbe, dotées d'une morphologie flexionnelle riche et d'une syntaxe flexible, posent des défis particuliers au TAL. Leur diversité aux niveaux lexical, morphosyntaxique et syntaxique se traduit par une dispersion des données : dans un corpus de taille standard, le nombre d'occurrences des phénomènes individuels reste bas, ce qui empêche les outils automatiques de les maîtriser. Ces langues sont donc souvent victimes d'un paradoxe : pour obtenir une bonne couverture des différents phénomènes qu'elles exhibent, elles doivent disposer de corpus plus larges que les langues à morphologie réduite. Or, la complexité de l'annotation exigée a un effet rédhibitoire sur la constitution de ressources et elles sont souvent relativement mal dotées en corpus annotés. Un indice en est la taille des corpus utilisés dans la campagne d'évaluation SPMRL 2013 (Seddah *et al.*, 2013). Ce fait est sans doute l'une des raisons pour lesquelles le serbe reste relativement peu doté de ressources et outils en TAL. Cependant, un autre facteur entre également en jeu : un manque de pratique du libre partage et de la diffusion des données et outils au sein de la communauté TAL serbe.

Au moment où nous avons entrepris ce projet, les seules ressources dédiées à cette langue librement diffusées comprenaient un corpus adapté à l'étiquetage morphosyntaxique (Krstev *et al.*, 2004), un corpus issu du Web doté d'annotations automatiques et par conséquent inadapté à l'entraînement des outils statistiques (Ljubešić et Klubička, 2014), un étiqueteur et lemmatiseur (Gesmundo et Samardžić, 2012) et un lexique morphosyntaxique (Krstev *et al.*, 2004). Cependant, de nombreuses autres

ressources sont citées dans les travaux existants sans être librement diffusées (Krstev et Vitas, 2005 ; Jakovljević *et al.*, 2014 ; Vitas et Krstev, 2004 ; Pavlović-Lažetić *et al.*, 2004 ; Krstev, 2008). Quant à l'analyse syntaxique, la seule tentative d'entraînement d'un analyseur syntaxique sur un corpus serbe a donné des résultats largement en dessous de l'état de l'art (LAS = 58 % et UAS = 66 %) (Jakovljević *et al.*, 2014), dus le plus probablement à la taille très limitée du corpus utilisé. À notre connaissance, ni le corpus d'entraînement ni les modèles d'analyse syntaxique développés n'ont été diffusés.

C'est dans ce cadre-là que nous avons posé les objectifs de ce travail : la création d'un corpus arboré pour le serbe, mais aussi de toute autre ressource qui pourrait faciliter le traitement automatique de cette langue, tel un lexique morphosyntaxique. Il faut néanmoins noter que cette situation s'est améliorée récemment grâce à la publication de plusieurs ressources, dont le lexique morphosyntaxique srLex (Ljubešić *et al.*, 2016) et un corpus arboré produit dans le cadre du projet UD, annoncé dans le travail de Samardžić *et al.* (2017) et publié en automne 2017.

Le croate, très proche du serbe, est mieux doté du point de vue du TAL. Avant la décomposition de l'ex-Yougoslavie, le serbo-croate était la langue officielle en Serbie, Croatie, Bosnie et au Monténégro. La création des États indépendants a mené à la proclamation des langues nationales. Leur statut est débattu depuis lors. Sans entrer dans des considérations socio-politiques complexes et sensibles, on peut résumer le rapport entre ces langues, à la suite de (Thomas, 1994), en disant que le serbe, le croate, le bosniaque et le monténégrin sont quasiment identiques aux niveaux phonologique, morphologique et syntaxique. Des différences plus importantes existent au niveau lexical, mais elles n'empêchent pas une compréhension mutuelle élevée des locuteurs sur le terrain. Parmi ces langues, c'est le croate qui est le mieux doté du point de vue du TAL (Agić *et al.*, 2013a ; Agić *et al.*, 2013b ; Agić et Ljubešić, 2014 ; Ljubešić *et al.*, 2016). Qui plus est, cette communauté pratique la libre diffusion de ressources et données. Nous nous sommes donc servis à plusieurs reprises de travaux effectués sur cette langue, ce qui sera détaillé dans la suite.

3. Constitution du corpus arboré serbe : méthode d'annotation adoptée

La création d'un corpus arboré est un processus complexe et coûteux. Avant d'entamer la constitution du corpus proprement dite, il est nécessaire de déterminer plusieurs aspects : il faut sélectionner le contenu à traiter, définir les annotations qui seront apportées au corpus, mettre en place des mécanismes pour assurer leur qualité, et optimiser le processus du point de vue du temps et de l'effort humain nécessaires. Dans cette section, nous présentons le corpus retenu (section 3.1), les principes qui ont guidé la création de notre corpus arboré (section 3.2) et la méthode qui nous a permis de les articuler (section 3.3).

3.1. *Corpus retenu*

Le contenu textuel utilisé dans ce projet provient de deux ouvrages littéraires serbes : *Bašta, pepeo* de D. Kiš et *Testament* de V. Stevanović (respectivements échantillons *basta* et *testament* dans le tableau 1)[3]. Une pratique plus courante consiste à utiliser des textes journalistiques (Marcus *et al.*, 1993 ; Abeillé *et al.*, 2003 ; Agić et Ljubešić, 2014), notamment parce que la question des droits limite souvent les possibilités de diffusion de textes littéraires. Or, nous avions déjà obtenu l'accord des ayants droit pour la diffusion non commerciale des textes concernés et avions procédé à leur étiquetage et, partiellement, à leur lemmatisation dans le cadre d'un travail antérieur (Miletic, 2013). Enfin, ce corpus littéraire apporte une diversification bienvenue aux corpus disponibles pour le croate et le serbe : SETimes.hr (Agić et Ljubešić, 2014) ainsi que les corpus UD pour le croate (Agić et Ljubešić, 2015) et le serbe (Samardžić *et al.*, 2017) sont basés sur des textes journalistiques.

La structure du corpus sélectionné et l'état de l'annotation au démarrage de la constitution du corpus arboré sont présentés dans le tableau 1.

Échantillon	*Tokens*	Étiquetage	Lemmatisation
Basta	55 783	Oui	Non
Testament	45 642	Oui	Oui
Total	101 425		

Tableau 1. *Structure et annotation préexistante de l'échantillon sélectionné*

3.2. *Principes d'annotation*

Quand il s'agit de l'annotation du corpus arboré, deux aspects principaux doivent être définis : la nature de l'annotation à apporter au corpus et les conditions dans lesquelles l'annotation manuelle se déroulera. Le premier dépend en partie de la nature de la langue traitée, mais aussi des exploitations envisagées du corpus et des contraintes temporelles du projet. L'objectif du deuxième est de garantir la qualité et l'efficacité du travail manuel.

Dans notre corpus arboré, nous avons mis en place une annotation en plusieurs couches. En effet, les exemples donnés dans la section 2.1.2 montrent que l'identification des fonctions syntaxiques en serbe repose fortement sur des traits morphosyntaxiques fins comme le cas, le nombre ou le genre. Nous avons donc effectué une annotation morphosyntaxique fine, qui inclut des traits utiles à l'analyse syntaxique. Au niveau de l'annotation syntaxique, nous avons adopté la syntaxe en dépendances (Tesnière, 1959 ; Mel'čuk, 1988). Au-delà du fait d'assurer une représentation

3. Kiš, Danilo. *Bašta, pepeo*, 2010. Podgorica : Narodna knjiga.
Stevanović, Vidosav. *Testament*, 1986. Beograd : SKZ.

aisée de structures discontinues (cf. exemple 2), ce cadre théorique devient le standard *de facto* en analyse syntaxique grâce aux campagnes d'évaluation CoNLL (Buchholz et Marsi, 2006 ; Nivre *et al.*, 2007) et au projet UD. Et comme nous nous attendions à un degré de dispersion des données important, nous avons inclus la lemmatisation pour réduire ce phénomène, suivant les travaux de Seddah *et al.* (2010) et de Le Roux *et al.* (2012). La contribution d'un lexique morphosyntaxique dans ce contexte ayant été mise en évidence (Hajič, 2000 ; Sagot, 2016), la création d'une telle ressource a également été incluse dans nos objectifs.

Quant aux conditions de l'annotation, nous avons conçu le processus de travail manuel de façon à garantir la validité de l'annotation tout en optimisant la vitesse de sa réalisation. Pour assurer la qualité et la cohérence des annotations manuelles, nous avons rédigé des guides d'annotation détaillés, qui ont été éprouvés *via* des évaluations de l'accord interannotateur. Ces techniques ont été utilisées dans la création de nombreux corpus arborés (Marcus *et al.*, 1993 ; Hajič, 2005 ; Brants, 2000).

Pour optimiser la rapidité du processus, nous avons eu recours à une préannotation automatique des données, dans la lignée de nombreux travaux qui ont démontré que cette méthode augmente la vitesse de traitement pour différents types d'annotation linguistique (Fort, 2012 ; Xue *et al.*, 2005 ; Tellier *et al.*, 2014). Dans la création des corpus arborés, cette méthode a été utilisée déjà lors de la constitution de PennTreebank (Marcus *et al.*, 1993) : ici, l'exploitation d'une préannotation automatique au niveau morphosyntaxique réduisait de moitié le temps d'annotation par rapport à l'annotation manuelle intégrale. Au niveau syntaxique, Chiou *et al.* (2001) notent que l'utilisation d'un analyseur syntaxique en constituants atteignant une précision de 82,87 % et un rappel de 81,42 % menait à une réduction du temps d'annotation de 50 %. Plus récemment, Skjærholt (2013) montre que la préannotation du norvégien avec un analyseur syntaxique d'une langue proche (le danois) permet de réduire le temps d'annotation de 50 %, et de 75 % lorsque l'analyseur est entraîné sur la même langue. Bien évidemment, l'utilisation de la préannotation entraîne le risque de biais : Fort et Sagot (2010) montrent que les erreurs d'un étiqueteur morphosyntaxique peuvent se propager dans les productions des annotateurs humains. Les auteurs concluent néanmoins que ce risque est justifié en vue des gains de temps conséquents. En témoignent également de nombreux autres projets de corpus arborés qui ont eu recours à cette approche (Hajič, 2005 ; Boguslavsky *et al.*, 2002 ; Abeillé *et al.*, 1998).

En l'absence d'outils déjà disponibles pour la préannotation du serbe, nous avons conçu une méthode itérative (section 3.3) fondée sur l'exploitation de ressources lexicales collaboratives (section 4.2) et l'adaptation de ressources disponibles pour une langue proche, le croate (section 4.3). La mise au point de ces différents volets de l'annotation s'appuie sur le principe général de l'annotation agile.

3.3. *Annotation agile basée sur un* bootstrapping *itératif multicouche*

L'organisation globale du travail est présentée dans la figure 1. Nous reprenons les quatre étapes de base identifiées par Fort (2012) : préparation (en bleu), précampagne (en jaune), campagne (en vert) et finalisation (en rouge). La phase de la campagne est plus complexe car elle est itérative et intègre des outils automatiques ; elle sera expliquée en détail dans la suite.

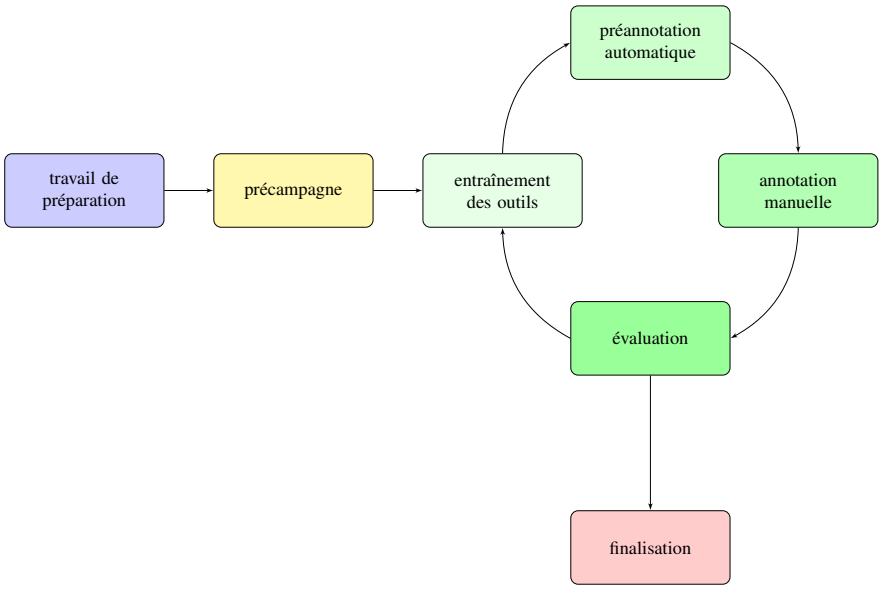


Figure 1. *Organisation du processus d'annotation*

Le travail de préparation correspond à la période de mise en place du matériel nécessaire à l'annotation du corpus. Concrètement, il s'agit de la sélection des outils automatiques à exploiter, du choix des textes qui composent le corpus, de la définition des jeux d'étiquettes, de la constitution des guides d'annotation et de leur première évaluation sur les données, ainsi que de la préparation des ressources externes (en particulier du lexique et des ressources d'entraînement initial pour les outils automatiques).

Le stade de la précampagne est dédié au recrutement et à la formation des annotateurs, qui leur permet de s'approprier les guides et les interfaces d'annotation.

L'organisation de la campagne est guidée par deux principes : l'agilité et l'utilisation d'outils automatiques. Le premier impose une organisation itérative du travail et introduit une étape d'évaluation à la fin de chaque cycle d'annotation manuelle. Le deuxième introduit deux étapes supplémentaires en début de chaque cycle : l'entraînement des outils et la préannotation automatique. Notons que ces deux étapes sont également exécutées itérativement par le recours au *bootstrapping* (v. *infra*). Lors du

premier passage par la boucle, l'entraînement des outils est effectué sur les ressources issues d'une phase minimale d'apprentissage constituées dans le stade du travail de préparation. Ces premiers modèles sont utilisés pour la préannotation du premier échantillon du corpus ; la préannotation automatique est corrigée manuellement, et l'échantillon nouvellement validé est rajouté aux ressources d'entraînement initiales. Lors du prochain passage par la boucle, les outils automatiques sont entraînés sur ces ressources augmentées, ce qui leur permet de s'améliorer à chaque itération, facilitant ainsi l'annotation manuelle. L'évaluation telle que nous la définissons diffère de ce qui est préconisé par Voormann et Gut (2008). Étant donné le temps nécessaire pour effectuer systématiquement l'annotation en double nécessaire à l'évaluation de l'accord interannotateur, nous n'intégrons pas celle-ci dans ce cycle, mais vérifions la qualité du travail des annotateurs à travers un contrôle ponctuel de la part d'un annotateur expérimenté. Cette étape contient également une séance de travail dédiée au retour d'expérience des annotateurs, qui porte notamment sur les guides d'annotation. Si les problèmes identifiés l'exigent, les guides sont modifiés en conséquence. Pour éviter les incohérences que ces modifications progressives peuvent introduire dans l'annotation, un travail d'harmonisation des annotations est réalisé dans la phase de finalisation.

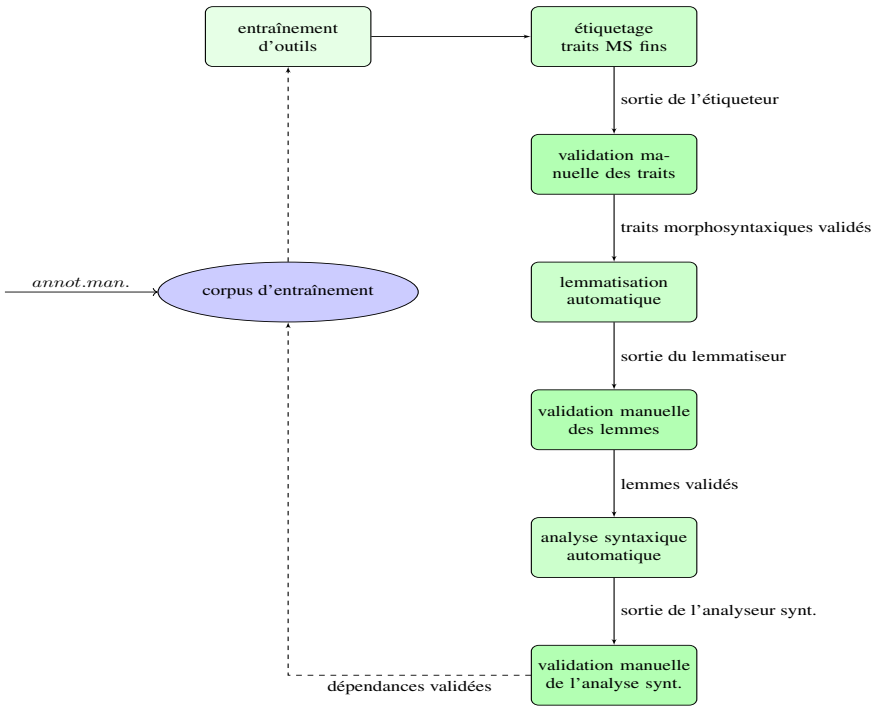


Figure 2. *Bootstrapping itératif pour une annotation multicouche*

La finalisation du corpus comprend donc le travail d'harmonisation des annotations mentionné ci-dessus. Elle porte également sur toutes les activités nécessaires à la diffusion du corpus : la conversion du corpus vers un format de diffusion standard, l'élaboration d'une documentation, la diffusion du corpus proprement dite.

Le schéma de la figure 1 correspond à l'annotation d'une seule couche d'informations. Notre corpus en exige trois. La figure 2 présente l'organisation d'une annotation multicouche, qui commence par un entraînement initial des trois outils dans l'étape de préparation. Dans le cadre de la campagne proprement dite, l'étiquetage morphosyntaxique, la lemmatisation et l'analyse syntaxique sont effectués en cascade. La sortie de chaque outil est corrigée manuellement avant d'être passée à l'outil suivant. Ainsi, chaque outil reçoit en entrée une annotation fiable, ce qui permet de maximiser ses performances et de faciliter par extension le travail des annotateurs humains. Précisons encore que l'ordre d'exécution des tâches est défini par les besoins des outils : l'analyseur syntaxique s'appuie typiquement sur les lemmes et les informations morphosyntaxiques, et le lemmatiseur exploite les catégories grammaticales, alors que l'étiqueteur n'exige pas de données supplémentaires.

4. Ressources optimisant l'annotation

Cette section est dédiée à une description du travail de préparation qui a porté sur la constitution des éléments matériels nécessaires à notre méthode.

4.1. *Création et évaluation des guides d'annotation*

Notre objectif étant ici de présenter la méthode globale, nous n'entrons pas dans les détails des schémas d'annotation adoptés pour la création du corpus arboré serbe, mais présentons plutôt les principes globaux sur lesquels ils reposent et le travail d'évaluation des guides d'annotation.

4.1.1. *Principes de constitution des jeux d'étiquettes et des schémas d'annotation*

Pour définir les jeux d'étiquettes morphosyntaxiques et syntaxiques, nous avons soumis les notions issues de la tradition grammaticale serbe à un examen détaillé théorique et empirique. Si les traitements existants ne se basaient pas sur des critères explicites, accessibles aussi bien à un annotateur humain qu'à un analyseur syntaxique, nous les avons modifiés de sorte à satisfaire cette exigence.

Au niveau morphosyntaxique, cela se traduit par l'élimination du jeu d'étiquettes d'un nombre de traits traditionnellement reconnus par les grammaires serbes et utilisés dans certains travaux de TAL (Krstev *et al.*, 2004)[4]. Il s'agit notamment de traits rele-

4. Le jeu d'étiquettes proposé dans ce travail compte 1 243 étiquettes. Une description détaillée est disponible à l'adresse suivante :`http://nl.ijs.si/ME/V4/msd/html/msd-sr.html`.

vant de distinctions qui ne sont pas utiles à l'identification des fonctions syntaxiques dans la phrase, comme l'aspect verbal, la définitude des adjectifs et l'opposition animé – non animé. Nous avons ainsi établi un jeu d'étiquettes raisonné de 1 042 étiquettes, basé sur les traits présentés dans le tableau 2.

Partie du discours	Traits encodés
Adjectif	Partie du discours, sous-catégorie, cas, nombre, genre, degré de comparaison
Nom	Partie du discours, sous-catégorie, cas, nombre, genre
Numéral	Partie du discours, sous-catégorie, cas, nombre, genre
Pronom	Partie du discours, sous-catégorie, cas, personne, nombre, genre
Verbe	Partie du discours, sous-catégorie, forme, personne, nombre, genre
Adverbe	Partie du discours, sous-catégorie, degré de comparaison
Conjonction	Partie du discours, sous-catégorie
Interjection	Partie du discours
Particule	Partie du discours
Préposition	Partie du discours

Tableau 2. *Traits morphosyntaxiques encodés dans le corpus arboré*

Au niveau syntaxique, les fonctions reconnues par les grammaires serbes (Stanojčić et Popović, 2012 ; Mrazović, 2009) ont été examinées à l'aide d'un ensemble de critères de distinction des relations syntaxiques de surface. Ces critères s'inspirent à la fois des fondements théoriques de la syntaxe en dépendances définis par Mel'čuk (1988) et du travail pratique de Burga *et al.* (2011) sur un corpus arboré espagnol. Dans ce travail, les auteurs établissent un inventaire de relations syntaxiques en espagnol en s'appuyant sur une liste de critères syntaxiques et morphologiques, comme les catégories grammaticales et les lemmes possibles du gouverneur et du dépendant, les traits de flexion du gouverneur et du dépendant, la présence de l'accord, la possibilité de pronominalisation par un clitique (dans le cas des dépendants verbaux), ainsi que les règles de linéarisation du gouverneur et du dépendant dans la phrase. Comme il s'agit de critères de surface, accessibles à un analyseur syntaxique, leur utilisation garantit non seulement une définition rigoureuse des étiquettes, mais aussi l'adaptation de ces dernières à l'analyse syntaxique. Nous avons donc adopté une démarche comparable.[5]

5. Une autre possibilité aurait consisté à adopter le jeu d'étiquettes du projet UD. Cependant, nous sommes en accord avec certaines critiques de ce schéma d'annotation proposées dans (Groß et Osborne, 2015) et nous avons opté pour une annotation basée sur le principe des têtes fonctionnelles et qui préserve un schéma particulier au serbe. Néanmoins, étant donné l'utilité du schéma UD pour un large éventail de recherches en TAL, la conversion du corpus existant vers ce format fait partie des perspectives du projet.

Cette démarche nous a permis d'identifier 48 étiquettes syntaxiques de base, complémentées par un traitement spécifique pour l'ellipse[6]. Dans le jeu, nous maintenons un noyau de fonctions de la grammaire serbe (les sujets, les objets, les prédicatifs). L'écart le plus important entre notre jeu d'étiquettes et la grammaire serbe relève de la distinction entre les arguments et les ajouts, que nous avons décidé d'ignorer, du fait des difficultés à la capter par des critères de surface. Nous avons opté pour une étiquette de dépendant sous-spécifiée. Une description détaillée du jeu d'étiquettes syntaxique est disponible dans le guide d'annotation diffusé avec le corpus[7].

4.1.2. *Mise au point des guides d'annotation*

Afin de garantir la complétude des guides d'annotation et leur adaptation à un travail sur corpus, nous les avons soumis à une évaluation de l'accord interannotateur. Nous avons calculé le taux d'accord pour l'annotation morphosyntaxique et l'annotation syntaxique, en utilisant comme mesure d'accord le *kappa* de Cohen (Cohen, 1960). Malgré des critiques quant à l'interprétation de ses différentes valeurs (Artstein et Poesio, 2008 ; Mathet *et al.*, 2012), le *kappa* de Cohen reste une mesure standard communément utilisée pour l'évaluation de l'accord interannotateur en corpus (Bhat et Sharma, 2012 ; Bond *et al.*, 2006 ; Agić et Merkler, 2013) et on considère en général que les valeurs supérieures à 0,90 représentent un accord satisfaisant.

Au niveau morphosyntaxique, l'évaluation a été effectuée par deux paires d'annotateurs. Chaque paire a eu un échantillon de 2 000 *tokens* à traiter. L'accord a été calculé aussi bien au niveau des étiquettes catégorielles (11 étiquettes instanciées) qu'au niveau des étiquettes morphosyntaxiques détaillées (205 étiquettes instanciées). Les résultats sont donnés dans le tableau 3. À titre d'illustration, le taux d'accord interannotateur au niveau morphosyntaxique dans le PennTreebank indiqué par Marcus *et al.* (1993) est de 96 % avec un jeu de 36 étiquettes, et celui dans le corpus arboré NEGRA est de 98,6 % (Brants, 2000) avec un jeu de 54 étiquettes. Notons néanmoins qu'il s'agit dans les deux cas d'un taux d'accord observé, qui n'est donc pas directement comparable à nos résultats.

Le degré de l'accord sur les étiquettes complètes est inférieur à celui qui concerne l'identification des catégories grammaticales. D'après le retour des annotateurs, ceci est en partie dû au nombre élevé de traits relatifs aux catégories grammaticales fléchies. Une analyse des matrices de confusion a montré que les distinctions les plus problématiques étaient celles entre les participes et les adjectifs déverbaux (cf. *izgu-*

6. Nous reprenons le traitement de l'ellipse mis en place dans le Prague Dependency Treebank (Hajič *et al.*, 1999, p. 204-221).

7. `https://github.com/aleksandra-miletic/serbian-nlp-resources/blob/master/ParCoTrain-Synt/`

bljen ‘perdu’) et celles entre les adverbes et les particules (cf. *jednostavno* ‘simplement’[8].). Le traitement de ces points a donc été clarifié dans le guide.

Annotation morphosyntaxique		
	Paire 1	Paire 2
kappa sur étiquettes complètes	0,90	0,91
kappa sur POS	0,96	0,97
Annotation syntaxique		
kappa sur fonctions syntaxiques	0,94	

Tableau 3. *kappa de Cohen à différents niveaux d’annotation*

Quant à l’annotation syntaxique, l’évaluation a été effectuée par une paire d’annotateurs sur un échantillon de 3 000 *tokens* (48 étiquettes instanciées) (cf. tableau 3). L’accord a été calculé en prenant en compte le rattachement étiqueté. À titre de comparaison, dans le corpus arboré des discussions Wikipédia FrWikiDisc, constitué par Urieli (2013), le *kappa* de Cohen était de 0,86 entre les deux annotateurs au niveau du rattachement étiqueté. Skjærholt (2013) indique un taux d’accord interannotateur observé de 95,3 % sur leur corpus arboré norvégien dans les mêmes conditions.

L’analyse de la matrice de confusion que nous avons effectuée a montré que c’est le traitement de différentes formes d’ellipse qui a généré le plus d’erreurs. Pour contrer cet effet, nous avons introduit des tests syntaxiques dans le guide d’annotation pour faciliter l’identification du gouverneur et du dépendant dans ces constructions.

Suite à ces modifications, les guides d’annotation ont été jugés suffisamment fiables pour être exploités dans le cadre de l’annotation manuelle.

4.2. *Création de ressources lexicales*

Comme mentionné dans la section 3.2, plusieurs travaux ont montré que l’utilisation d’un lexique externe peut faciliter l’analyse syntaxique d’une langue comme le serbe, car il permet de compléter le nombre de formes fléchies auxquelles un corpus d’apprentissage donne accès. Or, le seul lexique serbe librement disponible au début de ce projet était trop petit pour avoir un effet satisfaisant dans le cadre de l’analyse syntaxique (20 000 entrées seulement) (Krstev *et al.*, 2004). Pour assurer une couverture plus solide, nous avons constitué un lexique morphosyntaxique à partir du Wiktionnaire pour le serbo-croate, en nous inspirant des travaux de Sajous *et al.* (2013), Sagot (2014) et Sennrich et Kunz (2014), qui ont exploité la ressource libre du Wiktionnaire pour constituer des ressources électroniques dotées d’informations morphosyntaxiques.

8. Comparer l’adverbe extraprédicatif dans *Jednostavno, treba prestati* ‘Il faut simplement arrêter’ à l’adverbe intraprédicatif dans *Govori jednostavno* ‘Il parle simplement/d’une manière simple’.

Cette expérience, détaillée dans (Miletic, 2017), a abouti à la création de Wikimorph-sr, un lexique contenant 1 226 638 formes fléchies provenant de 117 445 lemmes différents, réparties en 3 066 214 triplets uniques <*forme fléchie, lemme, étiquette morphosyntaxique détaillée*>. Ce lexique est donc nettement mieux doté que le lexique existant du projet MultextEast (Krstev *et al.*, 2004) (20 000 entrées). Il est vrai que le contenu du lexique extrait n'est pas parfait : il présente un certain degré de surgénération des formes fléchies (notamment dans les paradigmes adjectivaux, qui disposent systématiquement des formes de comparatif et de superlatif, même pour les adjectifs relationnels), et ne contient pas du tout de catégories invariables, ce qui nous a amenés à effectuer des ajouts à partir de différentes ressources. Néanmoins, le lexique a été réalisé en trois semaines de travail, ce qui montre que cette approche est adaptée aux projets à durée limitée.

Ce travail a été effectué en 2015 ; en 2016, un lexique serbe de 5,3 millions d'entrées construit manuellement a été diffusé par Ljubešić *et al.* (2016). SrLex a été construit à partir des lexiques croate, serbe et bosniaque du logiciel de traduction automatique à base de règles Apertium (Forcada *et al.*, 2011), mais il a connu des extensions importantes dans le cadre d'une campagne de création manuelle d'entrées. Le lexique est librement diffusé[9], avec une licence qui autorise la redistribution. Nous avons fusionné les deux lexiques afin de maximiser leur utilité et avons ainsi obtenu une nouvelle ressource nommée ParCoLex. Ce troisième lexique contient au total 7 180 665 entrées uniques <*forme fléchie, lemme, étiquette morphosyntaxique détaillée*>, qui représentent 1 956 094 formes fléchies uniques provenant de 157 886 lemmes. Nous avons ensuite évalué la couverture des trois lexiques sur un échantillon de texte de 16 389 *tokens*, correspondant à 6 301 formes fléchies uniques. Les résultats, présentés dans le tableau 4, montrent que si SrLex a une couverture très largement supérieure à celle de Wikimorph-sr, la fusion des deux ressources offre un gain de 2,4 % sur les formes fléchies uniques, et de 4 % sur le nombre total d'occurrences des formes fléchies de l'échantillon. Ce lexique, aussi bien que Wikimorph-sr, est également diffusé avec le corpus arboré.

Lexique	Entrées	Lemmes	Couverture de l'autre lexique	Couverture échantillon	
				formes fléchies	occurrences
Wikimorph-sr	3 066 214	117 445	20,8 %	63,3 %	73,2 %
srLex	5 327 361	105 358	41,1 %	92,8 %	93,8 %
ParCoLex	7 180 665	157 886	NA	95,2 %	97,8 %

Tableau 4. *Tests de couverture avec les trois lexiques*

9. `http://nlp.ffzg.hr/resources/lexicons/srlex/`. Dernier accès : le 23 octobre 2017.

4.3. *Mise en œuvre des outils automatiques*

Notre choix d'outils de préannotation automatique a été guidé par deux critères : leurs performances sur le serbe ou sur une langue proche, et leur vitesse et ergonomie. Le premier critère s'explique par le fait qu'une préannotation de meilleure qualité facilite le travail des annotateurs humains, alors que le deuxième est dû à la nature itérative de notre méthode : comme elle prévoit plusieurs cycles d'entraînement et d'annotation, il est essentiel que les outils soient rapides et faciles à maîtriser.

À partir du travail d'Agić *et al.* (2013a), nous avons identifié l'étiqueteur HunPos (Halácsy *et al.*, 2007) et le lemmatiseur CST (Jongejan et Dalianis, 2009) comme outils offrant le meilleur compromis entre ces deux aspects. En ce qui concerne l'analyse syntaxique, les résultats signalés dans (Agić et Ljubešić, 2015) pointaient vers l'analyseur syntaxique Mate (Bohnet, 2010), basé sur un algorithme par graphes. Nous avons néanmoins préféré utiliser l'analyseur syntaxique Talismane (Urieli, 2013), basé sur un algorithme par transitions. Cet outil permet de définir avec précision l'exploitation de différents traits d'apprentissage (*tokens*, étiquettes POS, lemmes, informations morphosyntaxiques détaillées). Par ailleurs, il n'utilise pas les traits morphosyntaxiques désambiguïsés du corpus d'apprentissage, mais les puise plutôt dans un lexique externe en gardant toute l'ambiguïté rencontrée. Cette particularité est censée lui assurer une meilleure robustesse face au traitement d'un texte brut.

Une deuxième étape a consisté à assurer des ressources d'apprentissage pour les outils choisis et à les entraîner. Les travaux cités ci-dessus (Agić *et al.*, 2013a ; Agić et Ljubešić, 2015) montrent qu'il est possible de transposer un modèle entraîné sur le croate à des textes en serbe sans encourir d'importantes pertes de performances, et ceci aux trois niveaux d'annotation considérés. Comme les modèles de traitement développés par ces chercheurs sont librement disponibles, nous avons tâché de les exploiter pour la préannotation des échantillons de texte qui allaient constituer les ressources d'entraînement initiales pour notre méthode. Cette approche a été fructueuse pour l'étiquetage morphosyntaxique. Après avoir annoté notre échantillon avec le modèle de HunPos entraîné exclusivement sur le croate, nous avons constaté une exactitude moyenne du modèle de 77,95 %. Malgré la perte importante par rapport aux résultats rapportés dans (Agić *et al.*, 2013a) (où l'outil avait atteint une exactitude moyenne de 85 %), cette préannotation a facilité la correction manuelle de manière importante, permettant aux annotateurs humains de traiter 24 % de *tokens* en plus par rapport à une annotation manuelle intégrale. Cependant, le schéma d'annotation sur lequel a été entraîné le modèle croate n'est pas identique au nôtre, ce qui a entraîné des corrections supplémentaires pour adapter les traitements reproduits par l'étiqueteur à nos règles d'annotation. Ce sont ces modifications qui ont été jugées comme les plus chronophages par les annotateurs humains. Pour éviter ce type de corrections, dans l'étape suivante, le premier échantillon validé à l'aide du modèle croate a été utilisé pour ré-entraîner HunPos sur notre schéma d'annotation. Ce modèle, appris sur 20 000 *tokens*, a atteint une exactitude moyenne de 78,82 %, et la préannotation effectuée avec ce modèle a mené à une accélération du travail manuel de 60 % par rapport à l'annotation manuelle intégrale. La vitesse d'annotation moyenne dans ces conditions

était de 710 *tokens*/h. Pour comparaison, Marcus *et al.* (1993) indiquent une vitesse moyenne de 3000 *tokens*/h lors de la création du PennTreebank. Néanmoins, leur jeu d'étiquettes est beaucoup plus petit que le nôtre (36 étiquettes *vs* 1 042 étiquettes).

En revanche, cette démarche ne s'est pas avérée adaptée à la lemmatisation et à l'analyse syntaxique. En effet, le modèle croate de CST avait été entraîné sur des données étiquetées avec un jeu d'étiquettes morphosyntaxiques différent du nôtre ; par conséquent, ses performances sur nos données ont été compromises. Pour l'entraînement initial du lemmatiseur CST nous avons donc exploité un texte lemmatisé manuellement par Miletic (2013), que nous avons transformé en un lexique d'entraînement d'environ 20 000 entrées (10 000 lemmes différents). Malgré la taille restreinte de la ressource d'entraînement, ce premier modèle de CST a atteint une exactitude moyenne de 86,2 % et a permis une accélération de la lemmatisation manuelle de 41 % par rapport à la lemmatisation manuelle intégrale. Dans un deuxième temps et suite à la confection du lexique ParCoLex (cf. section 4.2), nous avons effectué un deuxième entraînement basé sur cette ressource. Le nouveau modèle a atteint une exactitude globale de 96,5 %, et la vitesse d'annotation manuelle a atteint 3 400 *tokens*/h, soit une accélération de 242 % par rapport à la lemmatisation manuelle intégrale.

Quant à l'analyse syntaxique, le modèle croate n'a pas pu être utilisé pour deux raisons : il a été entraîné sur un jeu d'étiquettes morphosyntaxiques différent et, qui plus est, il intégrait également un schéma d'annotation syntaxique très éloigné du nôtre. Nous avons donc jugé qu'une préannotation avec ce modèle n'était pas la manière la plus économe de procéder et avons favorisé une annotation manuelle. L'entraînement initial de Talismane a été effectué sur 40 000 *tokens* annotés à la main. Ce premier modèle a atteint une exactitude de 76,3 % en LAS et de 80,6 % en UAS. Ces scores ont été jugés satisfaisants, étant donné les résultats obtenus sur le serbe avec les analyseurs syntaxiques MST (LAS = 73,9 % et UAS = 80,6 %) (Agić *et al.*, 2013b) et Mate (LAS = 75,8 % et UAS = 82,4 %) (Agić et Ljubešić, 2015), notamment si l'on prend en compte le fait que les modèles en question ont été développés sur un corpus deux fois plus grand que notre échantillon (87 000 *tokens*).

5. Campagnes d'annotation manuelle

L'essentiel de l'annotation manuelle de notre corpus a été effectué dans le cadre de deux campagnes d'annotation avec des annotateurs étudiants serbophones. Les étudiants ont été majoritairement sélectionnés au département des études romanes à l'université de Belgrade à l'aide de questionnaires adaptés à la tâche à effectuer. Les campagnes d'annotation ont eu lieu à l'université de Toulouse-Jean-Jaurès. Avant d'entamer le travail d'annotation, les deux groupes d'annotateurs ont été formés à la fois sur les guides d'annotation et sur les interfaces d'annotation utilisées.

La première campagne a été dédiée à l'annotation morphosyntaxique et à la lemmatisation, effectuées en cascade et de manière itérative comme stipulé par notre méthode. En revanche, l'annotation syntaxique a été effectuée indépendamment durant

la deuxième campagne. Cette modification de notre méthode a été conditionnée par les compétences des annotateurs recrutés pour la première campagne. Les deux campagnes ont systématiquement fait appel aux outils de prétraitement présentés dans la section 4.3, entraînés selon la démarche de *bootstrapping* itératif. La tâche concrète des annotateurs consistait donc à corriger l'annotation produite par ces outils. Dans le cas de l'annotation syntaxique, nous avons exploité une fonctionnalité particulière de Talismane : cet analyseur syntaxique offre la possibilité d'accompagner chaque annotation du taux de probabilité qui lui est associé. Nous avons donc filtré la sortie de l'outil de sorte à ne retenir que les dépendances dont le taux de probabilité était supérieur à 0,85. Ainsi, les annotateurs disposaient d'une annotation partielle, mais fiable, ce qui a été jugé préférable par rapport à la possibilité de disposer d'une annotation complète, mais d'une qualité inférieure. Cette décision a été motivée par l'expérience de l'annotateur expert qui avait testé la préannotation intégrale dans l'étape de préparation. Cependant, nous n'avons pas fait d'expérience systématique sur ce point.

Les résultats des campagnes sont résumés dans le tableau 5. Comme certains annotateurs ont exprimé le souhait de poursuivre leur participation au projet, le travail d'annotation a été continué à distance (cf. les parties des deux campagnes effectuées à Belgrade). Précisons encore que l'annotation du corpus arboré a déjà été entamée, soit dans le cadre de travaux antérieurs, soit dans le cadre de l'initialisation des outils automatiques décrite dans la section 4.3. Le travail effectué par les annotateurs a donc mené à la complétion de l'annotation du corpus arboré.

Camp.	Tâche	Annotateurs	Durée	Endroit	Rendement	Pers.-heures
C1	morphosyntaxe	3 L3 + 1 M1	2 semaines	Toulouse	30 000 tok.	60 h
			6 semaines	Belgrade	30 000 tok.	50 h
	lemmatisation	3 L3 + 1 M1	2 semaines	Toulouse	35 000 tok.	25 h
C2	syntaxe	1 L3 + 1 M1	3 semaines	Toulouse	40 000 tok.	150 h
			6 semaines	Belgrade	20 000 tok.	60 h

Tableau 5. *Travail réalisé dans les campagnes d'annotation manuelle*

Nous pensons que ces résultats ont été favorisés par notre méthode globale : sans la préannotation automatique, la validation de différentes couches d'annotation aurait pris considérablement plus de temps. Par ailleurs, les annotateurs ont particulièrement apprécié le fait de pouvoir faire un retour direct de leurs expériences. Nous sommes d'avis que cette démarche a renforcé leur sentiment d'appartenance au projet, ce qui a garanti un niveau d'implication et de motivation élevé tout au long du projet.

6. Finalisation du corpus constitué

La dernière étape dans la création du corpus arboré a consisté en la finalisation et la diffusion du corpus constitué. Un annotateur expérimenté a été chargé d'harmoniser les annotations en accord avec la dernière version des guides d'annotation, établie

durant les campagnes d'annotation. Une fois ce travail finalisé, nous avons procédé à la préparation du corpus pour la diffusion.

Le corpus est diffusé dans le format CoNLL-X. Les champs contiennent les informations suivantes : l'ID du *token*, le *token*, le lemme, l'étiquette POS gros grain, l'étiquette POS plus spécifique, les traits morphosyntaxiques, l'ID du gouverneur du *token* et l'étiquette syntaxique du *token*. Le corpus a été divisé en trois sections : *train* (destinée à l'entraînement des analyseurs syntaxiques), *dev* (dédiée au paramétrage fin) et *test* (réservée à l'évaluation). Afin d'éviter tout biais relatif à la longueur des phrases dans différents segments du corpus, les phrases ont été réordonnées de manière aléatoire avant la segmentation du corpus en sections. Quelques informations sur les propriétés du corpus annoté sont données dans le tableau 6.

Les sections sont bien équilibrées au regard de plusieurs caractéristiques (longueur des phrases et profondeur de l'arbre syntaxique, représentation des parties du discours). Néanmoins, la répartition des étiquettes morphosyntaxiques détaillées est inégale. Ce fait pourrait se montrer problématique lors de l'utilisation du corpus pour l'entraînement des étiqueteurs morphosyntaxiques sur cette couche d'annotation.

Section	*Tokens*	Phrases	Formes fléchies	Lemmes	Étiq. POS	Étiq. détail.	Traits MS	Étiq. synt.	Long. phr.	Prof. ar.
all	101 029	3 861	22 739	11 251	16	679	165	67	27,16	6,98
train	80 869	3 116	19 598	10 120	16	643	159	67	26,95	6,98
test	10 162	367	4 033	2 802	15	432	134	60	28,69	7,11
dev	9 998	379	3 959	2 722	16	424	126	58	27,38	6,88

Tableau 6. *Corpus arboré constitué : statistiques de base. Long.phr. = longueur moyenne de phrase en* tokens*; Prof.ar. = valeur moyenne de la profondeur d'arbre maximale*

Le corpus est téléchargeable à partir de l'adresse suivante : `https://github.com/aleksandra-miletic/serbian-nlp-resources` sous la licence CC BY-NC-SA 3.0[10]. Ce dépôt contient également les différents modèles de traitement automatique développés durant ce projet pour l'étiquetage morphosyntaxique, la lemmatisation et l'analyse syntaxique, ainsi que la documentation des annotations apportées au corpus.

Ce corpus arboré s'est déjà montré utile dans plusieurs applications différentes. Notamment, il a permis d'entraîner un modèle d'analyse syntaxique avec Talismane qui atteint les scores de 87,48 % en LAS et de 91,22 % en UAS à partir d'un étiquetage morphosyntaxique manuel. Ceci représente un gain de 5,98 % en LAS et de 5,22 % en UAS par rapport aux meilleurs résultats préalables en analyse syntaxique du serbe, réalisés par Agić et Ljubešić (2015) avec l'analyseur syntaxique Mate. Il est vrai que certains aspects de ces expériences diffèrent de manière importante : le corpus d'entraînement utilisé par Agić et Ljubešić (2015) est journalistique alors que le nôtre est littéraire ; nos schémas d'annotation ne sont pas identiques (ils utilisent ceux du projet

10. `https://creativecommons.org/licenses/by-nc-sa/3.0/`

UD) ; différents types d'analyseurs syntaxiques ont été utilisés (Talismane est un analyseur syntaxique par transitions, alors que Mate est basé sur les graphes). Par conséquent, les résultats ne sont pas directement comparables. Néanmoins, les différences citées ne nous sont pas *a priori* favorables : les textes journalistiques sont en général considérés comme plus faciles que les textes littéraires, les analyseurs syntaxiques par graphes sont censés être plus adaptés aux langues à la morphologie flexionnelle riche, et les schémas d'annotation UD visent l'optimisation de l'analyse syntaxique indépendamment de la langue. Des évaluations plus directes sont tout de même nécessaires pour comprendre les effets de ces différents paramètres.

Notre corpus a également servi de base pour deux études en syntaxe théorique du serbe. La première version du corpus a été exploitée afin d'examiner les structures discontinues en serbe, à la fois dans une perspective monolingue et contrastive (Miletic et Urieli, 2017). La dernière version, présentée ici, a été utilisée dans une analyse approfondie de la position et de la structure du groupe adjectival gouverné par un nom (Miletic, 2018, ch. 10). Cette ressource s'est donc déjà montrée adaptée à la fois aux recherches en analyse syntaxique et en linguistique théorique.

7. Conclusions et perspectives

Dans cet article, nous avons présenté la démarche que nous avons utilisée pour constituer un nouveau corpus arboré pour le serbe, une langue peu dotée en corpus et outils du TAL. Ce corpus arboré s'accompagne également de lexiques morphosyntaxiques et de plusieurs modèles de traitement automatique. À la différence du corpus arboré serbe du projet UD, qui met en place l'annotation syntaxique indépendante de la langue préconisée par le projet, notre corpus est doté d'une annotation spécifique au serbe, qui établit une analyse plus fine, notamment au niveau des dépendants verbaux. Par ailleurs, le corpus UD est basé sur des textes journalistiques, alors que le nôtre relève du genre littéraire. Ce fait ouvre la voie aux expériences liées aux effets du genre, aussi bien en TAL qu'en linguistique serbe.

Nos résultats montrent que notre méthode, qui exploite au maximum les outils et ressources disponibles et qui accorde une attention particulière à l'organisation du travail des annotateurs, permet la réalisation aisée et relativement rapide d'une ressource complexe et polyvalente : la totalité du travail décrit ici a été effectuée entre novembre 2014 et avril 2018 dans le cadre d'un projet de thèse. Qui plus est, cette approche est basée sur des procédés généraux qui peuvent être appliqués à d'autres langues.

Tout d'abord, nous nous sommes servis de ressources existantes pour une langue proche : nous avons exploité avec succès un modèle d'étiquetage entraîné sur le croate pour faciliter la première phase d'étiquetage morphosyntaxique manuel de notre corpus. Grâce à la relation particulière entre le croate et le serbe décrite dans la section 2.2, nous avons pu effectuer cette manipulation sans faire appel à des techniques de traitement interlangue. Cependant, une approche comparable peut également être envisagée pour des langues plus éloignées. À titre d'illustration, le travail de Vergez-

Couret et Urieli (2015) montre que l'étiquetage de l'occitan bénéficie de l'ajout d'un grand corpus catalan à un corpus minimal de l'occitan lors de l'apprentissage.

Nous avons également fait appel à des ressources d'entraînement minimales : pour assurer un entraînement initial du lemmatiseur CST, nous avons exploité un lexique d'entraînement d'à peine 10 000 lemmes. Vu la richesse morphologique du serbe, ceci n'était pas prometteur. Néanmoins, malgré les performances moyennes de l'outil (86,2 % d'exactitude), la vitesse d'annotation manuelle a quasiment doublé.

Nous avons également exploité une ressource électronique créée de manière collaborative pour en extraire un premier lexique morphosyntaxique. En combinant le résultat de ce travail avec un autre lexique serbe (Ljubešić *et al.*, 2016), nous avons amélioré les résultats de la lemmatisation de manière significative. Cette approche est également transposable à toute langue dotée d'une ressource comparable au Wiktionary. À titre d'exemple, l'occitan et le picard en disposent.

Enfin, nous avons fait usage des particularités des outils utilisés pour la préannotation, et notamment de la capacité de l'analyseur syntaxique choisi à accompagner les étiquettes syntaxiques produites par les taux de probabilité associés. Ce fait nous a permis de trier sa sortie et de retenir une préannotation incomplète, mais fiable. Cette démarche est notamment importante lors d'une préannotation avec un outil relativement peu performant, autrement dit, durant les premières itérations de la démarche.

Dans l'avenir, nous poursuivrons ce travail dans deux directions principales. Premièrement, en ce qui concerne le corpus arboré créé, nous chercherons à raffiner certains points de l'annotation manuelle, et notamment le traitement des dépendants verbaux et de l'ellipse. Nous entreprendrons également la conversion du corpus vers le format UD, évoquée ci-dessus. Cette nouvelle annotation ne prendra cependant pas la place de l'annotation existante, mais sera ajoutée comme une couche d'information supplémentaire. Deuxièmement, nous chercherons à confirmer davantage la pertinence de notre méthode en la transposant sur l'occitan. Ce travail sera effectué dans le cadre du projet LINGUATEC (EFA227/16) « Développement de la coopération transfrontalière et du transfert de connaissance en technologies de la langue » du Programme de coopération territoriale Espagne-France-Andorre, POCTEFA, financé par le Fonds européen de développement régional. Cet effort démontrera le caractère transposable de notre méthode et sa capacité à favoriser la constitution des corpus arborés pour les langues peu dotées en général.

8. Bibliographie

Abeillé A., Clément L., Reyes R., « Talana annotated corpus : the first results », *Actes de The First Conference on Linguistic Resources*, Granada, 1998.

Abeillé A., Clément L., Toussenel F., « Building a treebank for French », *Treebanks*, Springer, p. 165-187, 2003.

Agić Ž., Ljubešić N., « The SETimes.HR Linguistically Annotated Corpus of Croatian », *Actes de Ninth International Conference on Language Resources and Evaluation (LREC2014)*,

European Language Resources Association (ELRA), Reykjavik, Iceland, p. 1725-1727, 2014.

Agić Ž., Ljubešić N., « Universal Dependencies for Croatian (that Work for Serbian, too) », *Actes de 5th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2015)*, p. 1-8, 2015.

Agić Ž., Merkler D., « Three syntactic formalisms for data-driven dependency parsing of Croatian », *Actes de 16th International Conference on Text, Speech and Dialogue*, Springer, Pilsen, Czech Republic, p. 560-567, 2013.

Agić Ž., Ljubešić N., Berović D., « Lemmatization and morphosyntactic tagging of Croatian and Serbian », *Actes de 4th Biennial International Workkshop on Balto-Slavic Natural Language Processing (BSNLP 2013)*, Sofia, Bulgaria, p. 48-57, 2013a.

Agić Ž., Merkler D., Berović D., « Parsing Croatian and Serbian by using Croatian dependency treebanks », *Actes de Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, p. 22-33, 2013b.

Alex B., Grover C., Shen R., Kabadjov M., « Agile corpus annotation in practice : An overview of manual and automatic annotation of CVs », *Actes de Fourth Linguistic Annotation Workshop*, Association for Computational Linguistics, Uppsala, Sweden, p. 29-37, 2010.

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

Bhat R. A., Sharma D. M., « A dependency treebank of Urdu and its evaluation », *Actes de Sixth Linguistic Annotation Workshop*, Association for Computational Linguistics, p. 157-165, 2012.

Boguslavsky I., Chardin I., Grigorieva S., Grigoriev N., Iomdin L. L., Kreidlin L., Frid N., « Development of a Dependency Treebank for Russian and its Possible Applications in NLP », *Actes de 3rd International Conference on Language Ressources and Evaluation (LREC2002)*, LREC, Las Palmas, Canary Islands, Spain, p. 852-856, 2002.

Bohnet B., « Very high accuracy and fast dependency parsing is not a contradiction », *Actes de 23rd International Conference on Computational Linguistics (COLING2010)*, Association for Computational Linguistics, Beijing, China, p. 89-97, 2010.

Bond F., Fujita S., Tanaka T., « The Hinoki syntactic and semantic treebank of Japanese », *Language Resources and Evaluation*, vol. 40, n° 3-4, p. 253-261, 2006.

Brants T., « Inter-annotator Agreement for a German Newspaper Corpus. », *Actes de 2nd International Conference on Language Ressources and Evaluation*, LREC, 2000.

Buchholz S., Marsi E., « CoNLL-X shared task on multilingual dependency parsing », *Actes de 10th Conference on Computational Natural Language Learning*, Association for Computational Linguistics, New York City, USA, p. 149-164, 2006.

Burga A., Mille S., Wanner L., « Looking behind the scenes of syntactic dependency corpus annotation : Towards a motivated annotation schema of surface-syntax in Spanish », *Actes de DepLing 2011*, p. 104-114, 2011.

Chiou F.-D., Chiang D., Palmer M., « Facilitating treebank annotation using a statistical parser », *Actes de First international conference on Human language technology research*, Association for Computational Linguistics, p. 1-4, 2001.

Cohen J., « A coefficient of agreement for nominal scales », *Educational and psychological measurement*, vol. 20, n° 1, p. 37-46, 1960.

Collins M., Ramshaw L., Hajič J., Tillmann C., « A statistical parser for Czech », *Actes de 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, Association for Computational Linguistics, p. 505-512, 1999.

Forcada M. L., Ginestí-Rosell M., Nordfalk J., O'Regan J., Ortiz-Rojas S., Pérez-Ortiz J. A., Sánchez-Martínez F., Ramírez-Sánchez G., Tyers F. M., « Apertium : a free/open-source platform for rule-based machine translation », *Machine translation*, vol. 25, n° 2, p. 127-144, 2011.

Fort K., Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus, PhD thesis, Université Paris-Nord-Paris XIII, 2012.

Fort K., *Collaborative Annotation for Reliable Natural Language Processing : Technical and Sociological Aspects*, John Wiley & Sons, 2016.

Fort K., Sagot B., « Influence of Pre-annotation on POS-tagged Corpus Development », *Actes de 4th Linguistic Annotation Workshop*, Association for Computational Linguistics, Uppsala, Sweden, p. 56-63, 2010.

Gesmundo A., Samardžić T., « Lemmatising Serbian as Category Tagging with Bidirectional Sequence Classification », *Actes de 8th Language Resources and Evaluation Conference (LREC 2012)*, p. 2103-2106, 2012.

Groß T., Osborne T., « The Dependency Status of Function Words : Auxiliaries », *Actes de 3rd International Conference on Dependency Linguistics (DepLing2015)*, Uppsala, Sweden, p. 111-120, 2015.

Hajič J., « Morphological tagging : Data vs. dictionaries », *Actes de 1st North American chapter of the Association for Computational Linguistics conference*, Association for Computational Linguistics, p. 94-101, 2000.

Hajič J., « Complex corpus annotation : The Prague dependency treebank », *Insight into Slovak and Czech Corpus Linguistics. Veda Bratislava*p. 54-73, 2005.

Hajič J., Panevová J., Buráňová E., Urešová Z., Bémová A., « Annotations at analytical level. Instructions for annotators », *UK MFF ÚFAL, Praha, Czech Republic. URL http ://ufal. mff. cuni. cz/pdt2. 0/doc/manuals/en/a-layer/pdf/a-man-en. pdf (2012-03-18)*, 1999.

Halácsy P., Kornai A., Oravecz C., « HunPos : an open source trigram tagger », *Actes de 45th annual meeting of the ACL on interactive poster and demonstration sessions*, Association for Computational Linguistics, Prague, Czech Republic, p. 209-212, 2007.

Hovy E., Lavid J., « Towards a 'science'of corpus annotation : a new methodological challenge for corpus linguistics », *International journal of translation*, vol. 22, n° 1, p. 13-36, 2010.

Jakovljević B., Kovačević A., Sečujski M., Marković M., « A Dependency Treebank for Serbian : Initial Experiments », *Actes de International Conference on Speech and Computer*, Springer, Novi Sad, Serbia, p. 42-49, 2014.

Jongejan B., Dalianis H., « Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike », *Actes de Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP.*, p. 145-153, 2009.

Keith B. (ed.), *Encyclopedia of language and linguistics*, 2006.

Krstev C., *Processing of Serbian. Automata, Texts and Electronic Dictionaries*, Faculty of Philology of the University of Belgrade, 2008.

Krstev C., Vitas D., « Corpus and Lexicon-Mutual Incompletness », *Actes de Corpus Linguistics Conference*, Birmingham, UK, p. 14-17, 2005.

Krstev C., Vitas D., Erjavec T., « MULTEXT-East resources for Serbian », *Actes de 7. mednarodne multikonference Informacijska druzba IS 2004 Jezikovne tehnologije 9-15 Oktober 2004, Ljubljana, Slovenija, 2004*, Erjavec, Tomaž and Zganec Gros, Jerneja, 2004.

Le Roux J., Sagot B., Seddah D., « Statistical parsing of Spanish and data driven lemmatization », *Actes d'ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages (SP-Sem-MRL 2012)*, p. 55-61, 2012.

Ljubešić N., Klubička F., « {bs, hr, sr} WaC–web corpora of Bosnian, Croatian and Serbian », *Actes de 9th Web as Corpus Workshop (WaC-9)*, Gothenburg, Sweden, p. 29-35, 2014.

Ljubešić N., Klubička F., Željko Agić, Jazbec I.-P., « New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian », *in* N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (eds), *Actes de Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, European Language Resources Association (ELRA), Paris, France, 2016.

Marcus M. P., Marcinkiewicz M. A., Santorini B., « Building a large annotated corpus of English : The Penn Treebank », *Computational linguistics*, vol. 19, n° 2, p. 313-330, 1993.

Marton Y., Habash N., Rambow O., « Dependency parsing of Modern Standard Arabic with lexical and inflectional features », *Computational Linguistics*, vol. 39, n° 1, p. 161-194, 2013.

Mathet Y., Widlöcher A., Fort K., François C., Galibert O., Grouin C., Kahn J., Rosset S., Zweigenbaum P., « Manual corpus annotation : Giving meaning to the evaluation metrics », *Actes d'International Conference on Computational Linguistics*, p. 809-818, 2012.

Mel'čuk I., *Dependency syntax : Theory and practice*, State University Press of New York, 1988.

Miletic A.*, « Annotation morphosyntaxique semi-automatique d'un corpus litéraire serbe »*, Master's thesis, Université Charles de Gaulle - Lille 3, 2013.

Miletic A., « Building a morphosyntactic lexicon for Serbian using Wiktionary », *Actes des Sixièmes Journées d'études Toulousaines (JéTou2017)*, Toulouse, France, p. 30-34, 2017.

Miletic A., Un treebank pour le serbe : constitution et exploitations, PhD thesis, Université de Toulouse Jean Jaurès, 2018.

Miletic A., Urieli A., « Non-projectivity in Serbian : Analysis of Formal and Linguistic Properties », *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling2017)*, Pisa, Italy, p. 135-144, 2017.

Mrazović P., *Gramatika srpskog jezika za strance*, Izdavačka knjižarnica Zorana Stojanovića, 2009.

Nivre J., Hall J., Kübler S., McDonald R., Nilsson J., Riedel S., Yuret D., « The CoNLL 2007 shared task on dependency parsing », *Actes de CoNLL shared task session of EMNLP-CoNLL*, sn, p. 915-932, 2007.

Pavlović-Lažetić G., Vitas D., Krstev C., « Towards full lexical recognition », *Text, Speech and Dialogue*, Springer, p. 179-186, 2004.

Pustejovsky J., Stubbs A., *Natural Language Annotation for Machine Learning : A guide to corpus-building for applications*, " O'Reilly Media, Inc.", 2012.

Sagot B., « DeLex, a freely-avaible, large-scale and linguistically grounded morphological lexicon for German », *Actes de 9th International Conference Language Resources and Evaluation (LREC2014)*, Reykjavik, Iceland, 2014.

Sagot B., « Etiquetage multilingue en parties du discours avec MElt », *Actes de la conférence conjointe JEP-TALN-RECITAL 2016*, Paris, France, 2016.

Sajous F., Hathout N., Calderone B., « Glàff, un gros lexique à tout faire du français », *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, Les Sables d'Olonne, France, p. 285-298, 2013.

Samardžić T., Starović M., Agić Ž., Ljubešić N., « Universal Dependencies for Serbian in Comparison with Croatian and Other Slavic Languages », *Actes de 6th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2017)*, Valencia, Spain, 2017.

Seddah D., Chrupała G., Çetinoğlu Ö., Van Genabith J., Candito M., « Lemmatization and lexicalized statistical parsing of morphologically rich languages : the case of French », *Actes de NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, Association for Computational Linguistics, p. 85-93, 2010.

Seddah D., Tsarfaty R., Kübler S., Candito M., Choi J., Farkas R., Foster J., Goenaga I., Gojenola K., Goldberg Y. *et al.*, « Overview of the SPMRL 2013 shared task : cross-framework evaluation of parsing morphologically rich languages », *Actes de Fourth Workshop on Statistical Parsing of Morphologically Rich Languages*, Association for Computational Linguistics, 2013.

Sennrich R., Kunz B., « Zmorge : A German Morphological Lexicon Extracted from Wiktionary », *in* N. C. C. Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (eds), *Actes de Ninth International Conference on Language Resources and Evaluation (LREC'14)*, European Language Resources Association (ELRA), Reykjavik, Iceland, 2014.

Skjærholt A., « Influence of preprocessing on dependency syntax annotation : speed and agreement », *Actes de 7th Linguistic Annotation Workshop and Interoperability with Discourse*, p. 28-32, 2013.

Stanojčić Ž., Popović L., *Gramatika srpskog jezika*, Zavod za udžbenike, 2012.

Tellier I., Eshkol-Taravella I., Dupont Y., Wang I., « Peut-on bien chunker avec de mauvaises étiquettes POS ? », *Actes de TALN 2014*, p. 125-136, 2014.

Tesnière L., « Eléments de syntaxe structurale », 1959.

Thomas P.-L., « Serbo-croate, serbe, croate..., bosniaque, monténégrin : une, deux..., trois, quatre langues ? », *Revue des études slaves*, vol. 66, n° 1, p. 237-259, 1994.

Urieli A., Robust French syntax analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit, PhD thesis, Université Toulouse le Mirail-Toulouse II, 2013.

Vergez-Couret M., Urieli A., « Analyse morphosyntaxique de l'occitan languedocien : l'amitié entre un petit languedocien et un gros catalan », *Actes de l'atelier TALARE 2015*, 2015.

Vitas D., Krstev C., « Intex and Slavonic morphology », *INTEX pour la linguistique et le traitement automatique des langues, Presses Universitaires de Franche-Comté*p. 19-33, 2004.

Voormann H., Gut U., « Agile corpus creation », *Corpus Linguistics and Linguistic Theory*, vol. 4, n° 2, p. 235-251, 2008.

Xue N., Xia F., Chiou F.-D., Palmer M., « The Penn Chinese TreeBank : Phrase structure annotation of a large corpus », *Natural language engineering*, vol. 11, n° 02, p. 207-238, 2005.

À l'écoute des locuteurs : production participative de ressources langagières pour des langues non standardisées

Alice Millour* — **Karën Fort****

Sorbonne Université, STIH - EA 4509, 28 rue Serpente, 75006 Paris, France
** alice.millour@etu.sorbonne-universite.fr, ** karen.fort@sorbonne-universite.fr*

RÉSUMÉ. *Les sciences participatives, et en particulier la production participative (*crowdsourcing*) bénévole, sont un moyen encore peu exploité de créer des ressources langagières pour les langues peu dotées dont suffisamment de locuteurs sont présents sur le Web. Nous présentons ici nos expériences concernant l'annotation en parties du discours pour des langues non standardisées, en l'occurrence l'alsacien et le créole guadeloupéen. Nous détaillons la méthodologie utilisée, montrons qu'elle est adaptable à plusieurs langues, puis nous présentons les résultats obtenus. L'analyse des limites de la plateforme d'origine nous a conduites à en développer une nouvelle, qui, outre l'annotation en parties du discours, permet la création de corpus bruts et d'un lexique de variantes alignées. Les plateformes créées, les ressources langagières, et les modèles de* taggers *entraînés sont librement disponibles.*

ABSTRACT. *Citizen science, in particular voluntary crowdsourcing, is still little experimented solution to produce language resources for less-resourced languages with enough connected speakers. We present here experiments we led on part-of-speech annotation for non standardized languages, namely Alsatian and Guadeloupean Creole. We detail the methodology we used and show that it is adaptable to other languages, then we present the results we obtained. An analysis of the limits of this platform led us to develop a new one, that allows the creation of raw corpora and part-of-speech annotations, and the construction of a multivariant lexicon. The created platforms, language resources and tagging models are all freely available.*

MOTS-CLÉS : *langues non standardisées, production participative, annotation en parties du discours.*

KEYWORDS: *non-standardized languages, crowdsourcing, part-of-speech annotation.*

TAL. Volume 59 – n°3/2018, pages 41 à 65

1. Pourquoi et comment créer des ressources pour des langues non standardisées

1.1. *Un enjeu culturel majeur*

Alors que les communications numériques connaissent un essor sans précédent et que l'accès aux technologies de communication moderne se démocratise[1], le monde numérique reste très peu représentatif des communautés linguistiques y ayant accès (Prado, 2012). Or, la diversité linguistique fait partie du patrimoine culturel à préserver, et la recherche en traitement automatique des langues (TAL), en permettant la saisie et la diffusion de contenus numériques, la présence en ligne de ressources linguistiques, ou encore le développement d'outils pédagogiques pour un nombre croissant de langues, peut participer à enrayer l'érosion à l'œuvre.

Les travaux portant sur de nouvelles langues présentent l'intérêt de confronter les chercheurs à des problématiques linguistiques nouvelles. Pour autant, la recherche en TAL ne concerne encore qu'une extrême minorité de langues : d'après Benjamin (2018), une majorité d'êtres humains a aujourd'hui pour langue maternelle l'une des 7 000 langues n'étant pas ou très peu considérées dans nos recherches. En cause, notamment, les politiques de financement et les opportunités professionnelles moindres découlant des recherches sur les langues peu dotées (Branco, 2018).

En outre, l'essor des communications virtuelles pose la question de l'intégration des langues non standardisées aux technologies du langage. Alors que l'UNESCO (Diki-Kidiri, 2007) préconise l'élaboration d'un certain nombre de ressources linguistiques (dont une orthographe et un système d'écriture, une grammaire écrite, un dictionnaire et une transcription phonétique) comme condition préalable à la présence pérenne d'une langue dans le cyber-espace, force est de constater que certaines de ces langues sont d'ores et déjà en usage sur Internet. L'absence de norme orthographique n'empêche pas l'utilisation de ces langues à l'écrit, en témoigne par exemple le cas de « l'*entre-soi* des groupes Facebook » (Rivron, 2012), qui favorise par exemple le dépassement de la gêne à écrire l'éton (langue de la région du Centre au Cameroun, autour de 250 000 locuteurs).

1.2. *Les sciences participatives, une solution encore peu exploitée*

La construction de ressources annotées de qualité par des linguistes est notoirement coûteuse[2]. De nombreuses langues, ne présentant pas un intérêt économique

1. Voir, par exemple, l'évolution de la couverture d'Internet donnée par Internet World Stats (`https://www.internetworldstats.com/stats.htm`), évaluée à 54,4 % de la population mondiale au 31 décembre 2017.

2. Voir par exemple (Böhmová *et al.*, 2001), l'une des rares publications donnant un coût approximatif pour une ressource langagière, le *Prague Dependency Treebank*, de l'ordre de 600 000 dollars.

immédiat, ou dont le nombre de locuteurs est faible, en sont par conséquent privées. Par ailleurs, ces locuteurs représentent un recours potentiel insuffisamment exploité pour la construction de ressources langagières. Notre hypothèse est qu'il est possible de pallier le manque de ressources langagières brutes et annotées en mettant à contribution les locuteurs *via* une plateforme de production participative adaptée intégrant des outils de TAL (y compris ceux créés à l'aide de ces ressources).

En outre, il apparaît que l'intervention des locuteurs dans la construction de ressources pour une langue non standardisée soit une condition nécessaire à la production de ressources de qualité représentatives des variétés de la langue.

1.3. *Une démarche expérimentale itérative*

L'hypothèse formulée pose un certain nombre de questions scientifiques : peut-on atteindre une qualité d'annotation suffisante de la part de locuteurs non familiers de la linguistique ? Quelle stratégie adopter pour assurer la qualité des ressources produites pour une tâche difficile pour les participants, susceptibles de commettre des erreurs ? Comment intégrer le TAL à la constitution de ressources de façon transparente pour le participant, tout en s'assurant qu'il ait conscience de l'impact de sa participation ? Enfin, comment optimiser, faciliter et rendre agréable une telle participation ?

Pour répondre à ces questions et évaluer notre hypothèse, nous avons développé deux plateformes de production participative permettant de recueillir des ressources de différentes natures. Ainsi, après avoir présenté la nécessité de construire des ressources annotées qui soient indépendantes de tout outil d'annotation, libres de droits, numérisées et accessibles, pour assurer la pérennité des avancées technologiques pour les langues peu dotées, nous présentons ici les deux expériences menées.

Notre première approche, décrite en partie 3 présente une méthodologie d'annotation participative en parties du discours. Si cette tâche est considérée comme résolue pour l'anglais et les langues dites « supercentrales » (Calvet, 2002)[3], de nombreuses langues, notamment l'alsacien et le créole guadeloupéen, ne bénéficient pas de corpus annotés de qualité ni d'outils performants. Nous détaillons dans la section 3.4 le processus d'instanciation de cette méthodologie pour ces deux langues de France non standardisées, ainsi que les résultats obtenus en termes de participation, de qualité des annotations et de performances des outils entraînés avec celles-ci. Cette première initiative de production participative pour deux langues peu dotées nous a amenées à développer une seconde méthodologie. Celle-ci place cette fois le locuteur au cœur de la production de corpus bruts représentatifs de la réalité des variétés existantes pour chaque langue avant même leur annotation, notamment grâce à la construction col-

3. Ce qui n'est vrai que dans une certaine mesure, comme le montre, entre autres, la quantité de travaux concernant l'adaptation au domaine, aux communications virtuelles (*computer-mediated communication*), ou aux contenus générés par les utilisateurs (*user-generated contents*), qui sont autant de catégories de productions langagières pouvant être considérées comme moins dotées au regard de certaines tâches.

laborative d'un lexique de variantes alignées. L'instanciation de cette méthodologie pour l'alsacien ainsi que les résultats préliminaires obtenus sont décrits en section 4.

2. État de l'art

Sont considérées comme peu dotées les langues qui, comparativement à d'autres, disposent de moins de ressources et outils favorisant leur intégration dans le monde numérique (voir notamment (Berment, 2004)). Cette appellation recouvre des réalités très variées : des langues ayant ou non le statut de langue officielle (par exemple, l'islandais et l'igbo), parlées par un nombre réduit ou important de locuteurs (par exemple l'inuktitut et le lao), présentant ou non une parenté avec une langue mieux dotée (par exemple l'occitan languedocien au regard du catalan, et l'arménien, isolé), etc. En résulte une grande diversité de caractéristiques linguistiques et de ressources, au sens large, disponibles pour chacune de ces langues.

Dans cette partie, nous présentons, dans un premier temps, les stratégies mises en place pour tirer parti de ces différents paramètres au regard de la tâche d'annotation en parties du discours. Puis, nous présentons les travaux existants quant à la production participative comme moyen de produire à bas coût des ressources de qualité.

2.1. *Annotation en parties du discours des langues peu dotées*

2.1.1. *Pallier le manque de ressources, approches existantes et limitations*

Les travaux existants quant à l'annotation en parties du discours présentent un éventail de stratégies visant à pallier le déficit de ressources annotées nécessaires au développement d'approches statistiques supervisées classiques. Elles diffèrent par la nature des ressources qu'elles requièrent. On compte notamment (i) les approches non supervisées tirant profit de la disponibilité de corpus parallèles permettant la projection d'annotations (Agić *et al.*, 2016), ou de la parenté de la langue considérée avec une langue mieux dotée (Hana *et al.*, 2004 ; Scherrer et Sagot, 2013 ; Bernhard *et al.*, 2018), (ii) les approches semi-supervisées telles que celle décrite par Garrette *et al.* (2013), intégrant par exemple des transducteurs à états finis pour analyser les mots inconnus, (iii) les approches faiblement supervisées, telles que l'utilisation du `Wiktionnaire` comme ressource complémentant un corpus annoté de taille réduite (Li *et al.*, 2012). Une solution pour limiter les coûts de développement est d'intégrer un lexique externe à l'entraînement d'un outil d'annotation supervisé de manière à augmenter la qualité d'annotation tout en limitant le coût de construction de la ressource (voir en particulier la figure 8.1 de (Sagot, 2018)).

Or, pour un grand nombre de langues, aucune de ces ressources n'est disponible en quantité suffisante. Par ailleurs, et quelle que soit la méthode employée, l'existence d'un corpus annoté est nécessaire, *a minima* comme référence pour évaluer les outils développés. Notons également que la libre disponibilité de ressources pérennes garantit la réutilisabilité de celles-ci, indépendamment des avancées technologiques.

2.1.2. *Le cas particulier des langues non standardisées*

Les langues non standardisées (ou *non canoniques* (Plank, 2016)) sont susceptibles de présenter des variations à tous les niveaux de l'analyse linguistique, de la phonétique à la sémantique. La question de leur intégration se pose dans quantité de cas dépassant celui des langues peu dotées, notamment celui des langues anciennes, par exemple le moyen allemand (Barteld, 2017), des contenus générés par les utilisateurs, comme `Wikipédia` (Krumm *et al.*, 2008), ou des communications médiées par ordinateur (Melero *et al.*, 2012). Des langues bien dotées, à l'instar du chinois mandarin (de Chine continentale, de Hong Kong et de Taïwan) (Tseng *et al.*, 2005) ou du portugais (brésilien et du Portugal) (Garcia *et al.*, 2014), sont également sujettes à cette variabilité. Or, à ce jour, les outils développés et évalués pour une langue donnée sont en réalité conçus de manière peu robuste à toutes formes de variations (Plank, 2016).

En ce qui concerne les langues peu dotées non standardisées, l'un des enjeux du respect de la diversité des variétés existantes est d'éviter de faire de la création de ressources et d'outils de TAL un vecteur non intentionnel de standardisation. La variabilité peut principalement être prise en compte de deux manières :

– soit en entraînant un outil pour chaque variété de langue considérée (*language adaptation*), ce qui implique, outre la nécessité de pouvoir identifier les variétés, une démultiplication du nombre de corpus d'entraînement nécessaires ;

– soit en normalisant les corpus (voir par exemple (Ljubešic *et al.*, 2016 ; Samardzic *et al.*, 2015), ou (Cox, 2010), pour une discussion sur la *rentabilité* de la normalisation). Cela suppose de définir une norme, et de connaître les variétés ainsi que les mécanismes de normalisation pour chacune d'elles. Est également envisagée l'utilisation de techniques de translittération (Pingali *et al.*, 2017), ou de dictionnaires de prononciation pour l'entraînement de modèles de transcription phonétique réduisant la variabilité scripturale (Steiblé et Bernhard, 2018).

Quelle que soit la méthode employée, celle-ci requiert soit une description de la variation existante, soit un large corpus permettant d'en inférer les motifs.

2.2. *Production participative de ressources langagières*

2.2.1. *Des productions participatives*

La production participative (*crowdsourcing*) s'est imposée depuis une dizaine d'années comme l'une des solutions aux freins que constitue le manque de moyens et de linguistes disponibles pour la construction de ressources langagières. Elle consiste à lancer un appel ouvert à participation (aujourd'hui principalement *via* le Web) pour faire réaliser une tâche, par des bénévoles (comme sur `Wikipédia`) ou en échange d'une (micro) rémunération (*microworking crowdsourcing*, comme sur `Amazon Mechanical Turk`[4]). Il existe bien entendu un continuum entre ces

4. Voir : `https://www.mturk.com/`.

deux extrêmes, avec des applications bénévoles offrant des « récompenses » variées, depuis le simple divertissement, à l'instar des jeux ayant un but comme JeuxDeMots (Lafourcade et Joubert, 2008) ou ZombiLingo (Guillaume *et al.*, 2016) jusqu'aux bons d'achat, comme sur Phrase Detectives (Chamberlain *et al.*, 2009).

Si le travail parcellisé à la Amazon Mechanical Turk permet d'accéder à une importante masse de travailleurs et de faire réaliser très rapidement des microtâches (HIT, *Human Intelligence Tasks*), ce type de plateforme ne permet pas de trouver plus facilement des experts pour certaines langues peu dotées (Callison-Burch et Dredze, 2010) et pose des problèmes éthiques et de qualité produite (Fort *et al.*, 2011). En effet, il est actuellement impossible, sur ces plateformes, de former les travailleurs à la tâche (il n'est possible que de les évaluer). Or, pour une tâche comme l'annotation en parties du discours, une formation est nécessaire : en témoignent les résultats obtenus par Hovy *et al.* (2014) pour l'annotation *via* CrowdFlower[5] de *tweets* en anglais (84 % d'exactitude), et de Jamatia et Das (2014) et Zaghouani et Dukes (2014) déplorant tous deux la faible qualité des annotations obtenues *via* Amazon Mechanical Turk, respectivement sur des *tweets* en hindi (moins de 60 % d'exactitude) et sur un chapitre du Coran (63,91 % d'exactitude).

Cette limitation n'existe pas dans les autres modes de production participative et de nombreuses plateformes imposent une formation (plus ou moins longue) préalable à la participation effective. C'est notamment le cas des Distributed Proofreaders[6], de Phrase Detectives ou de ZombiLingo. Les résultats ainsi obtenus en termes de qualité et de quantité de données produites sont tout à fait satisfaisants. Cependant, attirer et retenir les participants sur ces plateformes constituent un exercice complexe (Tuite, 2014), encore qualifiable d'« alchimie ».

2.2.2. *Productions participatives pour les langues peu dotées*

Les travaux concernant la production participative pour les langues peu dotées s'articulent selon différents axes. L'un concerne la production de données orales, dans un but de documentation des langues en danger, à l'instar des travaux de Bettinson et Bird (2017) et de Blachon *et al.* (2016) développant des outils de collecte de la parole *à l'attention des chercheurs*, ou de collecte d'informations sur la variation dialectale, dont par exemple (Leemann *et al.*, 2015). D'autres travaux utilisent la production participative pour recueillir des données géolinguistiques (voir par exemple (Avanzi et Stark, 2017) pour la variation du français, ou (Boula de Mareüil *et al.*, 2018) pour la construction d'un atlas sonore des langues régionales de France). À notre connaissance, sloWCrowd est la seule plateforme existante visant la production collaborative de ressources pour le traitement automatique des langues peu dotées. Si elle permet la validation de ressources annotées (par exemple, le sloWNet (Fišer *et al.*, 2014)) elle

5. Désormais Figure Eight, voir : https://www.figure-eight.com/.

6. Les Distributed Proofreaders corrigent les livres numérisés du Projet Gutenberg de mise à disposition de livres libres de droits : https://www.pgdp.net/c/.

est inadaptée à la résolution de tâches relativement complexes telles que l'annotation en parties du discours (Klubička et Ljubešić, 2014).

3. Expérimenter la production participative pour l'annotation en parties du discours

Notre première expérience de production participative a concerné l'annotation en parties du discours : à travers une plateforme dédiée, le participant annote des séquences de quatre phrases (figure 1) dont une provient d'un corpus de référence annoté par des linguistes et sert à l'évaluation du participant[7].

Figure 1. *Interface d'annotation en parties du discours pour l'alsacien*

3.1. *Une hypothèse de départ forte*

Les locuteurs des langues régionales de France sont très attachés à leur langue, à sa survie, voire à son développement. Notre hypothèse de départ a supposé qu'il n'était pas nécessaire de développer un jeu autour de la tâche pour motiver les locuteurs à y participer. Un tel développement étant en effet coûteux, nous avons privilégié la création d'une plateforme librement disponible légère, facile à maintenir et à adapter à d'autres langues. Nous avons par conséquent limité la ludification de la plateforme à un simple système de points permettant de classer les participants.

3.2. *Un processus cyclique intégrant un contrôle de la qualité*

3.2.1. *Préannotation et intégration continue des annotations*

Nous nous sommes inspirées de la méthodologie utilisée avec succès pour la syntaxe en dépendances du français dans `ZombiLingo` (Guillaume *et al.*, 2016), notam-

7. Le nombre de phrases a été choisi de manière à limiter la durée d'annotation d'une séquence (moins de dix minutes en moyenne), tout en assurant la collecte de suffisamment de données d'évaluation.

ment en intégrant i) des outils de préannotation, ii) une formation obligatoire pour les participants, et iii) une méthodologie d'évaluation continue des ressources produites.

La préannotation des corpus par deux outils intégrés à la plateforme est facultative mais permet de réduire la complexité de la tâche en ne proposant au participant que les étiquettes les plus probables. Si aucun outil, même imparfait, n'est disponible, une possibilité consiste à en entraîner un avec un corpus de taille très réduite dans un premier temps. Afin d'améliorer la qualité de la préannotation (ici, en parties du discours) au fur et à mesure de l'expérimentation, nous avons mis en place un processus vertueux (illustré par la figure 2), qui consiste à réentraîner régulièrement l'outil supervisé avec les annotations produites par les participants. Un deuxième outil de préannotation minimaliste peut être obtenu en tirant parti des caractéristiques linguistiques de la langue considérée (nous en donnons des exemples dans la section 3.4.4).

La double préannotation est utilisée de la manière suivante : lorsque les deux outils proposent la même étiquette, celle-ci est proposée en priorité au participant. Lorsqu'ils sont en désaccord, ce sont les deux étiquettes qui sont suggérées. Ce mécanisme permet d'accélérer la tâche de validation tout en laissant au participant la possibilité de choisir une autre étiquette dans la liste complète. L'outil supervisé étant réentraîné régulièrement, le taux d'accord entre les deux outils est amené à évoluer selon le processus vertueux lié aux performances croissantes de l'outil, facilitant à chaque itération la tâche au participant.

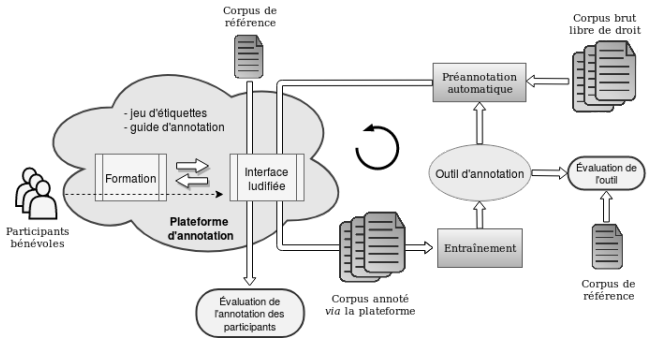


Figure 2. *Processus cyclique d'intégration continue des annotations produites*

3.2.2. *Une formation obligatoire*

Nous avons mis en place une phase de formation obligatoire pour tous les participants consistant à annoter intégralement quatre phrases issues d'un corpus de référence. Elle est conçue pour être la plus proche possible de la phase de production d'annotations, à ceci près que le participant ne peut valider une phrase que lorsque les annotations qu'il propose sont correctes. En cas d'erreur, le *token* mal annoté est mis en évidence mais l'étiquette attendue n'est pas divulguée. Cette première phase permet de confronter le participant aux difficultés de la tâche tout en le familiarisant aux

catégories existantes. Qu'il s'agisse de la phase de formation ou de production d'annotations, le participant à toujours accès à un guide d'annotation simplifié, organisé sous forme de listes d'exemples servant d'aide-mémoire pour chacune des catégories.

3.2.3. *Une évaluation continue des participants et des annotations*

La phase de production d'annotations est constituée d'une séquence de trois phrases à annoter issues du corpus brut, auxquelles s'ajoute une phrase issue du corpus de référence. Les $NbAnn_{Ref}$ annotations produites par un participant P sur cette phrase de référence permettent de calculer, à l'issue de chaque séquence, le score de confiance du participant : $Score_P = \frac{NbAnn_{Ref,Correctes}}{NbAnn_{Ref}}$. Ce score est ainsi mis à jour régulièrement et reporté sur toute annotation produite par le participant P sur un *token* T avec la catégorie C_i : $ScoreAnn_{T,P,C_i}$ vaut $Score_P$ au moment de l'annotation. Nous utilisons ce score de confiance pour filtrer les annotations de mauvaise qualité et pour identifier l'étiquette la plus probable parmi les éventuelles annotations concurrentes réalisées sur un *token* T par plusieurs participants. Nous déterminons ainsi pour chaque étiquette attribuée au *token* T un score de confiance $Score_{T,C_i}$ correspondant à la moyenne des scores des annotations Ann_{T,P_j,C_i} produites par différents participants : $Score_{T,C_i} = \frac{\sum_j ScoreAnn_{T,P_j,C_i}}{\sum_{i,j} ScoreAnn_{T,P_j,C_i}}$.

Nous choisissons enfin l'étiquette unique la plus probable pour chaque *token* : $C_T = \arg\max_i(Score_{T,C_i})$. Le corpus ainsi annoté est utilisé pour entraîner des *taggers*, utilisés à leur tour comme outils de préannotation dès lors que leurs performances dépassent l'outil précédent.

3.3. *Une méthodologie réplicable et des ressources produites librement disponibles*

Afin de proposer une méthodologie qui soit facilement réplicable[8], nous avons choisi d'utiliser le jeu d'étiquettes universel (Petrov *et al.*, 2012), contenant initialement 17 catégories (tableau 1), et facilement adaptable aux besoins spécifiques de chaque langue. Les modifications apportées à ce jeu d'étiquettes lors de l'instanciation de la plateforme sont présentées en partie 3.4.2.

Classes ouvertes	ADJ	ADV	INTJ	NOUN	PROPN	VERB		
Classes fermées	ADP	AUX	CCONJ	DET	NUM	PART	PRON	SCONJ
Autres	SYM	X	PUNCT					

Tableau 1. *Liste des étiquettes utilisées selon le classement de ses créateurs*

Par ailleurs, afin de garantir la disponibilité des ressources produites, la plateforme d'annotation est alimentée exclusivement de corpus libres de droits redistribuables.

8. C'est-à-dire que le processus peut être reproduit, sans nécessairement que les résultats le soient (reproductibilité) (Cohen *et al.*, 2018).

Pour des langues pouvant être peu dotées en termes même de ressources brutes, ce qui est souvent le cas des langues non standardisées, cela revient à suivre une démarche pragmatique aboutissant à la création de « corpus opportunistes » (McEnery et Hardie, 2011), représentant « [...] ni plus ni moins que les ressources qui ont pu être recueillies pour une tâche donnée. »[9].

Le code de la plateforme développée est quant à lui librement disponible sur `GitHub`[10] sous licence CeCILL v2.1[11]. La méthodologie décrite peut ainsi être adaptée facilement, la plateforme étant prête à être instanciée comme illustré dans la section 3.4.

3.4. *Une instanciation pour deux langues : l'alsacien et le créole guadeloupéen*

Nous avons mis en place deux instances de la plateforme décrite en section 3[12] : `Bisame`[13] (« *bisame* », ou « *bisanme* », « *bisàmme* », « *bisamme* », soit « ensemble » est employé dans l'expression « *Salü bisame !* », soit « Bonjour à tous ! ») pour l'alsacien, et `Krik`[14] (« *Krik* » est un terme intraduisible utilisé dans la tradition créole par les conteurs avant leur prise de parole) pour le créole guadeloupéen[15].

3.4.1. *Deux langues de France aux profils très différents*

L'alsacien est un terme générique pour le continuum de sous-systèmes dialectaux germaniques (Malherbe, 1983) parlé en Alsace et dans une partie de la Moselle. Le bas alémanique, variété principale de l'alsacien, est lui-même divisé en deux sous-ensembles : le bas alémanique du nord (NV) et du sud (SV). On trouve à Strasbourg une variété du bas alémanique du nord légèrement teintée de francique (STRV). En dépit du déclin de la transmission familiale, une étude comptabilise 550 000 locuteurs en 2004 (Barre et Vanderschelden, 2004). Le créole guadeloupéen, à base lexicale française et africaine, compte pour sa part environ 600 000 locuteurs (400 000 en Guadeloupe, 200 000 ailleurs dans le monde (Colot et Ludwig, 2013)). L'atlas des langues en danger établi par l'UNESCO[16] ne donne pas le degré de vitalité de l'alsacien pris

9. "*[...] nothing more nor less than the data that it was possible to gather for a specific task.*" (McEnery et Hardie, 2011).

10. Voir : `https://github.com/alicemillour/Bisame`.

11. Voir : `http://www.cecill.info/`.

12. Ces plateformes ont fait l'objet de publications spécifiques (Millour et Fort, 2018a ; Millour et Fort, 2018b), nous les présentons ici en regard, avec de nouveaux éléments d'analyse, notamment une enquête sur les participants.

13. Voir : `http://bisame.paris-sorbonne.fr`.

14. Voir : `http://krik.paris-sorbonne.fr`.

15. Alsacien et créole guadeloupéen évoluant dans un contexte de diglossie avec le français, c'est la langue que nous avons utilisée pour nos interfaces. Outre la dimension pratique de ce choix, cela nous permet également de ne pas avoir à préférer une variété dialectale ou scripturale à une autre, évitant ainsi d'exclure une partie des locuteurs.

16. Voir : `http://www.unesco.org/languages-atlas/fr/atlasmap.html`.

isolément, mais donne celui du groupe des « langues alémaniques », comprenant également le souabe et le haut valaisan. Ce groupe, au nombre de locuteurs de l'ordre du million, est classé comme vulnérable par l'UNESCO. Le créole guadeloupéen, plus dynamique, est pour sa part absent de l'atlas.

Aucune de ces langues n'a été lissée par l'usage d'une forme normative écrite, bien que des initiatives de graphies unifiées existent, notamment l'orthographe ORTHAL (Crévenat-Werner et Zeidler, 2008) pour l'alsacien, et celle du GEREC-F (Groupe d'études et de recherches en espace créolophone et francophone) (Ludwig *et al.*, 1990) modifiée plus tard par Bernabé (2001) et coexistant avec le système introduit par Hazaël-Massieux (2000) pour le créole guadeloupéen. En résulte une variabilité scripturale qui s'additionne à la variabilité dialectale, multipliant les graphies existantes pour un élément de lexique donné. Enfin, les deux langues coexistant avec le français, les éventuels trous lexicaux sont parfois remplis par des termes français.

3.4.2. *Tokénisation et mises à jour des jeux d'étiquettes*

La tokénisation de langues non standardisées représente une gageure, dans la mesure où l'ensemble des pratiques scripturales n'est pas connu en amont de la conception du tokéniseur. Dans le cadre de nos expériences, nous avons utilisé un script `Python`, initialement développé pour l'alsacien (Bernhard *et al.*, 2017) que nous avons adapté au créole guadeloupéen. Dans les deux langues, le script a été mis à jour tout au long de l'expérimentation, de nouvelles formes orthographiques remettant en cause nos règles de tokénisation apparaissant au fur et à mesure que nos corpus augmentaient en taille.

Parallèlement à ces ajustements du tokéniseur, le jeu d'étiquettes (voir la section 3.3) a été complété pour les deux langues, afin de faciliter l'annotation sans affecter la bonne lisibilité des textes proposés. En particulier, nous avons dû effectuer un certain nombre de choix arbitraires assurant la bonne intelligibilité des séquences proposées pour les participants, *a priori* non experts.

Dans le cas de l'alsacien, nous avons ajouté la catégorie ADP+DET pour les contractions n'étant pas automatiquement séparées, par exemple *am*, contraction de *an* et *dem* (« au »). Dans le cas du créole guadeloupéen, la tokénisation, par exemple, de la contraction *k'ay*, regroupant *ka*, particule du présent, et *ay*, (3^{e} personne du singulier du verbe « avoir »), sous forme de deux *tokens k'* et *ay*, rendait la lecture et la compréhension difficiles pour les locuteurs. Ces considérations nous ont amenées à ajouter la catégorie PART+VERB. Pour les mêmes raisons, les *tokens* contenant des pronoms tels que *ba'y* (« pour lui/elle »), *trapé'y* (« l'attraper »), ou *sa'w* (contraction de *sa* (« ce/cette ») et *ou* (« tu »), littéralement « ce que tu ») nous ont amenées à l'ajout des catégories ADP+PRON, VERB+PRON et PRON+PRON.

3.4.3. *Collecte des corpus bruts et construction d'une référence minimale*

Les corpus bruts alimentant la plateforme ont été construits suivant la méthodologie décrite dans la section 3.3. Nous avons recueilli pour ces deux langues l'ensemble

des corpus libres de droits et accessibles à notre connaissance. Ceux-ci proviennent par conséquent de sources hétérogènes, telles que les projets de la Wikisphère (Wikipédia, Wiktionnaire, incubateurs Wikimédia), des textes non soumis au droit d'auteur produits par les organismes locaux de promotion de la langue, notamment l'OLCA[17], ou gracieusement fournis par des participants, ou de bases de données telles COCOON[18] qui contient des transcriptions de conversations en créole guadeloupéen[19]. Les contenus et tailles de ces corpus sont détaillés dans le tableau 2. Les variétés des corpus alsaciens (section 3.4.1) sont données en indice. Dans le cas de COCOON, la taille est donnée en nombre de groupes de souffle transcrits, pas de phrases.

	Nom	Nb. phrases (Nb. *tokens*)	Source
Alsacien	$T_{\texttt{gsw,SV}}$	267 (5 110)	Wikipédia
	$T_{\texttt{gsw,STRV}}$	66 (1 768)	Nouvelle
Créole guadeloupéen	$T_{\texttt{gcf,Wiki}}$	74 (873)	Wikisphère
	$T_{\texttt{gcf,COCOON}}$	1 080 (9 175)	COCOON

Tableau 2. *Corpus collectés pour l'alsacien et le créole guadeloupéen*

Les corpus de référence, annotés manuellement par des chercheuses du laboratoire LiLPa de Strasbourg pour l'alsacien et par une étudiante guadeloupéenne, deux expertes de l'annotation et un dialectologue créolophone pour le créole guadeloupéen, contiennent respectivement 102 et 100 phrases. Leurs contenus sont détaillés dans le tableau 3.

	Nom	Nb. phrases (Nb. *tokens*)	Source
Alsacien	E_{SV}	47 (875)	Wikipédia
	$E_{\text{NV,1}}$	26 (362)	Pièce de théâtre
	$E_{\text{NV,2}}$	29 (231)	Recettes
Créole guadeloupéen	$E_{\texttt{gcf,Wiki}}$	17 (238)	Wikisphère
	$E_{\texttt{gcf,COCOON}}$	83 (1 385)	COCOON

Tableau 3. *Corpus de référence annotés par des experts linguistes*

3.4.4. *Outils de préannotation*

Comme décrit dans la section 3.2.1, nous avons intégré deux outils de préannotation à chacune des instances développées. Dans le cas de l'alsacien, nous avons utilisé le Stanford POS Tagger (Toutanova *et al.*, 2003) pour l'allemand, selon la méthodologie définie par Bernhard et Ligozat (2013), ainsi que MElt (Denis et Sagot, 2010),

17. Office pour la langue et la culture d'Alsace, voir https://www.olcalsace.org/.
18. Collection de corpus oraux numériques, voir https://cocoon.huma-num.fr/.
19. Voir par exemple, sous licence CC BY-NC-SA : https://cocoon.huma-num.fr/exist/crdo/meta/crdo-GCF_1022.

entraîné au fur et à mesure de la croissance du corpus d'entraînement annoté *via* la plateforme. Dans le cas du créole guadeloupéen, aucun outil d'annotation n'étant disponible à notre connaissance, nous avons développé un script `Python` tirant parti de la faible flexion du créole guadeloupéen et de l'importante fréquence absolue des *tokens* les plus fréquents : par exemple, la particule *ka* représente 4,6 % du corpus brut, le pronom *an* (« je »), 3,6 %, le verbe *sé* (verbe « être », sous ses formes infinitive et conjuguées), 2,8 %, etc. Nous avons extrait du corpus de référence une liste des 100 couples *token*-étiquette non ambigus les plus fréquents que nous avons utilisés pour annoter le corpus brut. Cette liste n'est pas représentative des mots les plus fréquents en créole guadeloupéen, mais nous a néanmoins permis d'annoter 37 % du corpus.

Nous savons que la préannotation introduit un biais (Fort et Sagot, 2010), auquel les utilisateurs les moins formés sont les plus sensibles (Dandapat *et al.*, 2009). Il est donc probable que celle-ci a un impact sur nos participants. Nous avons néanmoins observé, dans le cas de l'alsacien, que si les outils proposent la même étiquette dans 50 % des cas en moyenne, celle-ci est rejetée par les participants dans 12 % des cas.

3.5. *Résultats obtenus et discussion*

3.5.1. *Participation*

	Alsacien	Créole guadeloupéen
Nombre d'inscrits	208	35
Participants ayant finalisé la phase d'entraînement	75	17
Participants ayant produit des annotations	47	11
Jours d'annotation	109	9
Nombre d'annotations produites	24 588	1 205
Taille du corpus annoté (*tokens*)	7 973	933
Qualité des annotations produites (F-mesure)	0,93	0,87

Tableau 4. *Participation sur les deux plateformes*

La participation sur les plateformes est détaillée dans le tableau 4. L'écart entre les deux plateformes (plus de 200 participants pour l'alsacien et 35 pour le créole guadeloupéen) s'explique à notre avis par la différence d'énergie déployée à communiquer sur chacune des instances : la publicité de la plateforme `Bisame` a été réalisée à travers des communications sur les groupes Facebook de locuteurs, par le bouche-à-oreille, grâce au relais d'organisations comme le FILAL[20] ou d'entreprises telles que la Marque Alsace[21], par le biais d'une chronique diffusée sur France Bleu Elsass[22], et *via* la page Facebook du projet `Bisame`[23]. La plateforme `Krik` n'a pas bénéficié

20. Fonds international pour la langue alsacienne, voir `https://filalsace.net/`.
21. Voir : `http://www.marque-alsace.fr/`.
22. Voir : `https://www.francebleu.fr/elsass`.
23. Voir : `https://www.facebook.com/bisame.elsass/`.

d'un tel effort de communication, nos contacts étant moindres et l'expérimentation ayant été rapidement interrompue. Nous avons en effet constaté que le corpus que nous avions à disposition pour le créole guadeloupéen rendait l'annotation trop difficile, voire impossible, notamment parce que le jeu d'étiquettes utilisé n'était pas adapté à l'annotation de l'oral. En effet, le corpus du créole est constitué en majorité de transcriptions et est découpé en groupes de souffle. Ces séquences à annoter se sont révélées inutilisables en l'état car inintelligibles du fait de la présence de nombreux achoppements et de structures syntaxiques incomplètes. N'ayant pas à notre disposition d'autres corpus libres de droits pour le créole guadeloupéen, et un autre projet incluant la production bénévole de corpus bruts étant en cours de développement (voir la section 4), nous avons mis cette instance en pause.

Nous avons également observé, et cela est valable pour les deux instances, qu'environ 40 % des participants ne produisent aucune annotation après avoir finalisé la phase de formation. On peut supposer que la durée de la formation (huit minutes en moyenne) ainsi que la nature de la tâche, difficile et rébarbative sont la cause de cette démotivation. Par ailleurs, bien que l'application soit conçue pour être utilisable sur téléphone mobile, son inconfort d'utilisation a été évoqué par plusieurs participants comme un facteur de découragement.

Nous avons réalisé une enquête auprès des participants de la plateforme `Bisame`[24] afin de recueillir des informations sur leurs genres, âges, niveaux d'études, et langues maternelles. Cette enquête montre qu'il existe une marge de progression quant à la participation, notamment des femmes. En effet, sur les 22 participants ayant répondu à l'enquête, 77 % sont des hommes, ce qui va à l'encontre des observations de Chamberlain *et al.* (2013), qui montrent que les femmes sont davantage enclines à participer à ce genre d'interface ludifiée. Par ailleurs, près de 30 % des répondants ont pour langue maternelle le français et non l'alsacien et 36 % déclarent avoir au-delà de 60 ans, une majorité ayant entre 21 et 40 ans. Enfin, les participants ont un niveau d'études élevé, 60 % d'entre eux ayant atteint au moins le niveau BAC + 4, ce qui participe à expliquer la bonne qualité des annotations obtenues.

3.5.2. *Ressources produites*

Les difficultés liées à la nature du corpus se ressentent dans la qualité des annotations produites sur la plateforme `Krik` (voir le tableau 4) : celles-ci atteignent une exactitude de 87 %, très en deçà de ce que nous observons sur `Bisame` (93 %). Notons tout de même que ces résultats sont supérieurs à celui obtenu pour une tâche semblable réalisée par travail parcellisé avec `CrowdFlower` : 84 % d'exactitude pour de l'annotation de *tweets* en anglais (Hovy *et al.*, 2014).

Pour comprendre la source des erreurs commises par les participants, nous avons corrigé manuellement le corpus annoté *via* la plateforme `Krik`. Outre les difficultés liées au corpus, évoquées plus haut, cette analyse nous a permis de révéler des li-

24. Nous ne présentons pas les résultats de l'enquête menée pour la plateforme `Krik`, celle-ci ayant reçu trop peu de réponses pour être exploitable.

mitations de notre tokéniseur dues à l'apparition d'habitudes scripturales inconnues. Par exemple, la forme séparée *anba la* (« en dessous ») génère deux *tokens*, qui, lorsqu'ils ne sont pas suivis d'un nom commun ne peuvent pas être annotés séparément. Ils doivent par conséquent être regroupés sous la forme *anba_la* pour être annotés comme adverbe. Ces cas de figure ont été intégrés progressivement au script de tokénisation grâce à de nouvelles règles. Nous avons par ailleurs analysé les erreurs commises par les participants de manière à en identifier les motifs récurrents. Par exemple, de nombreuses confusions existent entre les catégories ADJ et VERB dans le cas de l'alsacien, ou le cas de *té*, pouvant désigner le verbe « être » ou la particule désignant le passé en créole guadeloupéen. Ont ainsi été mis en évidence les cas les plus intrinsèquement ambigus requérant une vigilance particulière et devant être intégrés à la phase de formation ainsi qu'au guide d'annotation mis à disposition.

3.5.3. *Outils entraînés*

Nous avons utilisé les ressources produites pour entraîner le *tagger* `MElt`, et observé deux types de difficultés propres à chacune des plateformes[25]. Dans le cas du créole guadeloupéen, nous avons dû compenser la mauvaise qualité des annotations recueillies : la correction manuelle du corpus annoté a permis d'augmenter de 10 % les performances de l'outil entraîné, passant de 76 à 84 % d'exactitude. Dans le cas de l'alsacien, l'entraînement de différentes instances de `MElt` avec différents sous-corpus correspondant à des variétés spécifiques de l'alsacien a permis de mettre en avant la nécessité de prendre en compte ces variétés. L'intégration de deux lexiques préexistants (environ 40 000 entrées, décrits dans (Millour et Fort, 2018b)) à l'entraînement de `MElt`, permet d'améliorer les performances de l'outil sur les mots inconnus de près de 30 % en moyenne, mais ne suffit pas à produire une couverture lexicale suffisante pour compenser les variabilités dialectale et lexicale.

Ce cas est illustré dans le tableau 5 : avec trois variétés présentes en tout (SV et STRV dans le corpus d'entraînement, SV et NV dans le corpus d'évaluation), on constate que les meilleures performances sont obtenues lorsque le corpus d'entraînement et le corpus de test appartiennent à la même variété SV (83,7 %). D'autre part, on observe que les performances sur l'ensemble des corpus d'évaluation E_{SV}+$E_{\mathrm{NV},1}$+$E_{\mathrm{NV},2}$ augmentent très faiblement malgré l'ajout du corpus T_{STRV} (1 768 *tokens*). Nous observons que l'augmentation globale des performances (+ 1,2 point) peut se faire au détriment de la qualité d'annotation sur certains corpus d'évaluation pris séparément, ici E_{SV} (- 1,4 point). En effet, malgré une augmentation de 30 % de la taille du corpus d'entraînement, le pourcentage de mots inconnus est stable pour ce corpus et certains *tokens*, en quantité insuffisante ici pour conclure sur la nature de la baisse de performances, se retrouvent mal annotés. Ce phénomène met en avant l'importance de la prise en compte des différentes variétés de l'alsacien dans le développement d'outils d'annotation performants.

25. Une analyse complète des résultats et des performances des outils entraînés a été présentée dans (Millour et Fort, 2018a ; Millour et Fort, 2018b).

	E_{Sv}	$E_{\text{Nv},1}$	$E_{\text{Nv},2}$	E_{Sv}+$E_{\text{Nv},1}$+$E_{\text{Nv},2}$
T_{Sv}	**83,7**	78,7	71,3	77,9
Unk. *tokens*	40 %	65 %	62 %	52 %
$T_{\text{Sv}} + T_{\text{STRV}}$	82,3	**82,8**	**71,8**	**79,1**
Unk. *tokens*	40 %	37 %	61 %	47 %

Tableau 5. *Exactitude des outils entraînés pour l'alsacien*

3.5.4. *Enseignements tirés*

En ce qui concerne la participation et les quantités limitées de ressources produites, deux éléments d'analyse nous semblent importants à considérer.

D'une part, notre hypothèse de départ concernant la motivation des locuteurs à développer des outils pour leur langue s'est révélée insuffisante. En effet, si contribuer à la création de ressources langagières pour leur langue motive certains à venir participer, ils ne restent pas (Millour et Fort, 2017). La même observation a été faite dans un cadre extrême, celui d'une mission humanitaire visant à traduire des SMS pour aider les rescapés du tremblement de terre à Haïti en 2010 (Munro, 2013) : les volontaires se sont épuisés (dans tous les sens du terme) au bout de quelques semaines. Cette constatation rejoint les résultats de l'enquête concernant `ZombiLingo` : si certains jouent pour « aider les scientifiques », ceux qui reviennent et participent le plus le font pour le jeu (Fort *et al.*, 2017). Il est donc nécessaire d'ajouter des éléments ludiques pour favoriser la rétention des participants.

Par ailleurs, les variétés scripturales et dialectales inhérentes aux langues non standardisées posent plusieurs types de problèmes. D'une part, il est plus facile pour un locuteur (qu'il soit linguiste ou non) d'annoter la variété d'une langue qui lui est la plus familière. Dans le cadre d'un projet de production participative tel que le nôtre, il apparaît que proposer différentes variétés est indispensable, afin de ne pas perdre de contributeurs. En témoignent les commentaires reçus par mail et *via* le formulaire de contact mis en place sur la plateforme :

> « *J'ai dernièrement envoyé le lien vers le site à des membres de ma famille d'origine alsacienne... ils me demandent maintenant s'il faut contribuer en haut-rhinois ou en bas-rhinois... auriez-vous une idée ?* »
>
> « *C'est de l'alsacien haut-rhinois, pas toujours facile pour les gens du Bas-Rhin ! On a fait ce qu'on a pu.* »

La création d'une instance pour le créole guadeloupéen a montré la facilité technique de l'adaptation de la plateforme. Il est néanmoins indispensable de disposer de corpus bruts de taille et de qualité suffisantes pour permettre une annotation de qualité. En outre, la collaboration avec des locuteurs influents au sein de la communauté linguistique est apparue comme un facteur important de réussite pour ce type de projet participatif.

Enfin, l'absence de prise en compte de la variation peut conduire à une stagnation voire à une dégradation des performances de l'outil entraîné. Or, pour certaines langues peu dotées, en particulier non standardisées, il existe peu, voire pas, de corpus disponibles, ni de description des mécanismes de variabilité à l'œuvre. Le processus de collecte de corpus ne peut donc se faire sans l'intervention des locuteurs, ceux-ci étant les seuls à même de produire des corpus écrits qui soient *représentatifs* des variétés scripturales et dialectales en usage. Les ressources produites *via* les plateformes `Bisame` et `Krik` sont disponibles sous différentes licences libres fonctions des corpus bruts correspondants[26]. Les modèles de *taggers* entraînés sont également disponibles, sous licence CC BY-NC-SA[27].

4. `Recettes de Grammaire` : une plateforme autonome de création de ressources pour les langues non standardisées

La première plateforme a permis de valider une méthodologie tout en montrant ses limites. La suivante tire les enseignements de cette expérience, notamment en permettant la production de corpus (dans un premier temps, des recettes de cuisine) et la construction d'un lexique de variantes alignées. Nous avons également largement renforcé la fluidité du processus d'annotation et la ludification des tâches proposées.

4.1. *Produire et annoter des corpus dans différentes variétés de la langue*

La plateforme `Recettes de Grammaire` répond à trois objectifs : i) faire produire du corpus brut sous forme de recettes de cuisine et d'anecdotes, ii) faire corriger les annotations produites par un outil état de l'art et iii) faire produire des variantes scripturales et dialectales pour les mots des recettes. L'ajout de recettes, d'anecdotes et de commentaires est réalisé par le biais d'une interface classique inspirée de sites de cuisine existants. Outre son intérêt linguistique, la plateforme a donc également un rôle culturel, puisqu'elle permet de produire une base de données de recettes qui, on peut l'espérer, seront typiques de la région.

4.2. *Fluidifier le processus d'annotation*

Afin d'encourager un participant ayant ajouté une recette à l'annoter dans la foulée, nous avons mis en place le cheminement suivant :

1) le participant ajoute une recette ;

2) cette recette est préannotée à la volée par un outil état de l'art et le résultat de cette annotation est montré au participant ;

26. Les articles `Wikipédia` sous licence CC BY-SA, les autres textes sous licence CC BY-NC-SA.
27. Voir : `https://bisame.paris-sorbonne.fr/downloads` (pour l'alsacien) `https://krik.paris-sorbonne.fr/downloads` (pour le créole guadeloupéen).

3) si le participant accepte de corriger ces préannotations, il est renvoyé vers l'interface correspondante.

L'annotation, qui est en réalité une correction de la préannotation fournie par l'outil, se fait de manière séquentielle *par catégorie* : en cliquant sur une catégorie, le participant fait apparaître les préannotations correspondantes qu'il peut valider ou rejeter. La phase d'annotation est structurée en différentes étapes correspondant à trois niveaux de difficulté des catégories établis grâce à l'étude des annotations produites par les participants (voir 3.5.1) :

– *facile*, ayant obtenu plus de 0,95 de F-mesure (pour l'alsacien, NOUN et DET) ;

– *intermédiaire*, ayant obtenu une F-mesure comprise entre 0,90 et 0,95 (pour l'alsacien, ADP+DET, NUM, INTJ, PROPN, VERB et SYM) ;

– *difficile*, pour les catégories restantes (pour l'alsacien, ADJ, ADV, ADP, AUX, CONJ, PRON, PART, SCONJ et X).

Les trois étapes de l'annotation sont les suivantes :

1) dans un premier temps, seules les catégories *faciles* sont visibles. Cette étape est destinée à mettre en confiance le participant quant à sa capacité à participer à l'amélioration des performances de l'outil entraîné ;

2) une fois les catégories *faciles* annotées, la liste complète des étiquettes apparaît à droite du texte saisi (voir la figure 3). Les catégories hachurées sont les catégories *difficiles* qui requièrent une formation préalable. La formation à une catégorie consiste à présenter au participant une séquence de phrases dont certains *tokens* ont été annotés manuellement avec la catégorie donnée. Le participant doit valider les préannotations correctes et rejeter celles qui sont erronées pour pouvoir valider sa formation. Les catégories blanches sont les *intermédiaires* entre lesquelles le participant peut naviguer librement, les grisées sont celles qui n'ont pas été utilisées lors de la préannotation ;

3) une fois toutes les étiquettes issues de la préannotation examinées par le participant, il lui reste à annoter les *tokens* dont l'étiquette a été rejetée, en choisissant une étiquette parmi la liste complète.

Ce découpage permet de diminuer la complexité de la tâche par rapport à une annotation séquentielle. La formation est en outre plus ciblée. Enfin, le participant peut naviguer entre les catégories, accéder au guide d'annotation pour chacune d'entre elles sous la forme d'un menu déroulant, et corriger ses annotations s'il le souhaite.

4.3. *Évaluer les participants, une gageure*

La nature des textes annotés ne permet pas d'introduire des phrases de référence sur lesquelles évaluer les participants. Il nous a donc été impossible de reproduire la méthodologie d'évaluation, pourtant efficace, présentée dans la section 3.2.3. Néanmoins, nous pouvons évaluer les participants en introduisant des préannotations volontairement erronées, pour un jeu de *tokens* connus et non ambigus. Par exemple,

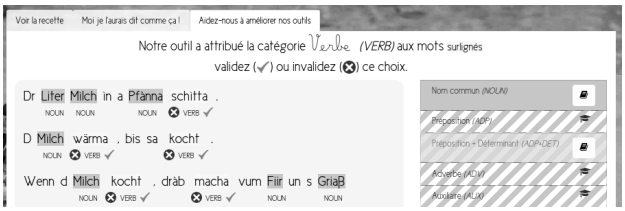


Figure 3. *Extrait de l'interface d'annotation pour la catégorie* VERB

la préannotation du mot « avec » *mìt*/ADV doit être corrigée en *mìt*/ADP. Les performances du participant sur ces *tokens* de test définissent son niveau de confiance. Cette méthode, sans doute moins efficace que la précédente, fera l'objet d'une attention particulière pour être améliorée au besoin.

4.4. *Annoter sa variété : mise en place d'un outil d'édition*

Afin de construire un lexique de variantes alignées nous renseignant sur les mécanismes de variations à l'œuvre et pouvant être intégré à l'entraînement d'outils supervisés, nous avons mis en place l'interface « Moi je l'aurais dit comme ça ! » (voir la figure 4). Elle permet à tout participant d'ajouter une variété scripturale ou dialectale d'un mot d'une recette. Les participants ayant la possibilité de placer leur lieu d'apprentissage de l'alsacien sur une carte de l'Alsace découpée en cinq aires dialectales, nous envisageons par ailleurs d'utiliser cette fonctionnalité pour leur proposer les contenus existants dans la variété qu'ils préfèrent afin de faciliter leur participation (voir 3.5.4).

Figure 4. *Ajout de la variante* Kugelhof *pour le mot* Kugelhopf

4.5. *Améliorer la ludification*

La plateforme `Recettes de Grammaire` est plus stylisée et personnalisée que les plateformes précédentes notamment grâce à un profil plus complet, un accès aux profils publics des participants et à leurs recettes, et à la possibilité de commenter les contenus et d'interagir entre participants au sein de la plateforme. Aux fonctionnalités existantes (nombre de points et classement des joueurs), nous avons ajouté un ensemble de badges récompensant l'activité des participants sur la plateforme. Outre les badges s'accumulant à mesure que le participant ajoute recettes, anecdotes, variantes et annotations, nous avons introduit des badges de compétence obtenus à l'issue des formations réalisées sur les catégories *difficiles*. Le système de points a également été rendu plus interactif : deux participants proposant la même étiquette pour un *token* donné voient leurs points doubler. Le participant ayant annoté en premier en est averti *via* une fenêtre *pop-up* à la connexion suivante.

5. Discussions et conclusion

Les résultats de nos premières expériences de production participative d'annotations en parties du discours pour des langues peu dotées sont encourageants. Elles ont en effet abouti à la création d'une plateforme *open source*, d'un corpus annoté de 7 973 *tokens* pour l'alsacien, présentant une F-mesure de 0,93, et d'un premier corpus de référence annoté de 2 439 *tokens* pour le créole guadeloupéen. Les outils entraînés grâce à ces corpus atteignent une exactitude allant de 71 % à 84 % en fonction de la variété du corpus d'évaluation pour l'alsacien, et de 84 % pour le créole guadeloupéen. La mise en ligne des ressources annotées sur ORTOLANG[28], afin d'en assurer la pérennisation, est en cours. La démarche de production participative dans laquelle nous nous inscrivons est un processus cyclique dans lequel le dialogue avec le locuteur fait partie intégrante du développement. Les enseignements tirés de ces premières expériences nous ont donc conduites à développer la plateforme `Recettes de Grammaire` intégrant la collecte de corpus et de lexique de variantes alignées.

Cette nouvelle plateforme, lancée en juin 2018, a permis de recueillir à ce jour de premiers résultats encourageants (9 recettes, 515 annotations, 110 variantes dialectales et scripturales), qui sont en cours de traitement.

La participation à nos plateformes est très respectable si on la compare à d'autres portant sur des langues plus importantes en termes de nombre de locuteurs. Ainsi, la première version de `Phrase Detectives` (pour l'anglais) a attiré 2 000 joueurs en 32 mois (Chamberlain *et al.*, 2013), ce qui est proportionnellement bien inférieur pour une langue parlée par environ 350 millions de personnes[29]. Cependant, l'investissement des communautés concernées doit et peut être amélioré. Il apparaît notamment que les locuteurs militant pour la survie de leur langue n'ont pas conscience du fac-

28. Accessible ici : `https://www.ortolang.fr/`.

29. Selon `Wikipédia` : `https://en.wikipedia.org/wiki/English-speaking_world`.

teur aggravant que constitue l'absence de ressources numériques et d'outils adaptés. Nous pensons que les projets de production participative tels que le nôtre doivent ainsi également servir de vecteur de sensibilisation. Il nous revient donc d'atteindre la communauté des jeunes apprenants, par exemple en facilitant l'utilisation sur mobile des plateformes. Nous espérons enfin que la diversification des tâches présentée en section 4 équilibrera la répartition des participants, en termes de genre notamment.

Par ailleurs, la mobilisation des participants étant coûteuse quant à la communication qu'elle requiert, il est nécessaire que leur investissement soit exploité au mieux. L'utilisation de la préannotation et la fluidité d'utilisation des plateformes sont deux moyens mis en place permettant d'optimiser leur participation. Par ailleurs, l'étude des erreurs qu'ils commettent nous renseigne sur le niveau de difficulté *pour les annotateurs* et nous permet d'améliorer notre méthodologie à cet égard, notamment en proposant des formations ciblées. La comparaison des performances de `MElt` avec celle d'autres *taggers* supervisés nous permettra d'identifier où se situe la difficulté *pour les outils*, et de l'intégrer à la conception de notre plateforme.

Si l'interface `Recettes de Grammaire` ne rencontrait pas le succès espéré, la modularité des composants développés nous permettrait de modifier facilement le mécanisme d'incitation à la production de corpus bruts. Enfin, le code source de la nouvelle plateforme est librement disponible[30], ce qui permettra à tous ceux qui le souhaitent de l'améliorer et de l'adapter à leurs propres besoins.

Remerciements

Nous remercions vivement les participants du projet PLURAL (production ludique de ressources linguistiques pour le TAL) Langues et numérique 2018 (DGLFLF) : Delphine Bernhard, Bruno Guillaume et André Thibault, ainsi que les contributeurs des projets `Bisame` et `Krik` et l'OLCA pour son soutien. Nous remercions également la première relectrice pour ses nombreuses remarques et ses conseils.

6. Bibliographie

Agić Ž., Johannsen A., Plank B., Martínez H. A., Schluter N., Søgaard A., « Multilingual projection for parsing truly low-resource languages », *Transactions of the Association for Computational Linguistics*, vol. 4, p. 301-312, 2016.

Avanzi M., Stark E., « A crowdsourcing approach to the description of regional variation in French object clitic clusters », *Belgian Journal of Linguistics*, 2017.

Barre C., Vanderschelden M., *L'enquête "étude de l'histoire familiale" de 1999 - Résultats détaillés*, INSEE, Paris, 2004.

Barteld F., « Detecting spelling variants in non-standard texts », *Actes de Student Research Workshop (EACL 2017)*, Valence, Espagne, mai, 2017.

30. Voir : `https://github.com/alicemillour/Bisame/tree/recipes`.

Benjamin M., « Hard Numbers : Language Exclusion in Computational Linguistics and Natural Language Processing », *Actes de 11th International Conference on Language Resources and Evaluation (LREC'18)*, Miyazaki, Japon, mai, 2018.

Berment V., Méthodes pour informatiser les langues et les groupes de langues "peu dotées", Thèse, Université Joseph-Fourier - Grenoble I, mai, 2004.

Bernabé J., *La graphie créole*, Ibis Rouge edn, Guides du CAPES de Créole, 2001.

Bernhard D., Ligozat A.-L., « Es esch fàscht wie Ditsch, oder net ? Étiquetage morphosyntaxique de l'alsacien en passant par l'allemand », *Actes de TALARE (Traitement Automatique des Langues Régionales de France et d'Europe) (TALN'13)*, Les Sables d'Olonne, France, p. 209-220, juin, 2013.

Bernhard D., Ligozat A.-L., MARTIN F., Bras M., Magistry P., Vergez-Couret M., Steible L., Erhart P., Hathout N., Huck D., Rey C., Reynés P., Rosset S., Sibille J., Lavergne T., « Corpora with Part-of-Speech Annotations for Three Regional Languages of France : Alsatian, Occitan and Picard », *Actes de 11th edition of the Language Resources and Evaluation Conference (LREC'18)*, Miyazaki, Japon, mai, 2018.

Bernhard D., Todirascu A., MARTIN F., Erhart P., Steible L., Huck D., Rey C., « Problèmes de tokénisation pour deux langues régionales de France, l'alsacien et le picard », *Actes de DiLiTAL (Diversité Linguistique et TAL) (TALN'17)*, Orléans, France, juin, 2017.

Bettinson M., Bird S., « Developing a suite of mobile applications for collaborative language documentation », *Actes de 2nd Workshop on Computational Methods for Endangered Languages*, Honolulu, Hawaï, p. 156-164, mars, 2017.

Blachon D., Gauthier E., Besacier L., Kouarata G.-N., Adda-Decker M., Rialland A., « Parallel Speech Collection for Under-resourced Language Studies Using the Lig-Aikuma Mobile Device App », *Procedia Computer Science*, vol. 81, p. 61-66, 2016.

Böhmová A., Hajič J., Hajičová E., Hladká B., « The Prague Dependency Treebank : Three-Level Annotation Scenario », *in* A. Abeillé (ed.), *Treebanks : Building and Using Syntactically Annotated Corpora*, Kluwer Academic Publishers, 2001.

Boula de Mareüil P., Rilliard A., Frédéric V., « A Speaking Atlas of the Regional Languages of France », *Actes de 11th International Conference on Language Resources and Evaluation (LREC'18)*, Miyazaki, Japon, mai, 2018.

Branco A., « We Are Depleting Our Research Subject as We Are Investigating It : In Language Technology, more Replication and Diversity Are Needed », *Actes de 11th International Conference on Language Resources and Evaluation (LREC'18)*, Miyazaki, Japon, mai, 2018.

Callison-Burch C., Dredze M., « Creating speech and language data with Amazon's Mechanical Turk », *Actes de Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT '10) de NAACL HLT 2010*, Association for Computational Linguistics, Los Angeles, CA, États-Unis, juin, 2010.

Calvet L.-J., *Le marché aux langues : Essai de politologie linguistique sur la mondialisation*, Plon, 2002.

Chamberlain J., Fort K., Kruschwitz U., Lafourcade M., Poesio M., « Using Games to Create Language Resources : Successes and Limitations of the Approach », *in* I. Gurevych, J. Kim (eds), *The People's Web Meets NLP*, Theory and Applications of Natural Language Processing, Springer Berlin Heidelberg, p. 3-44, 2013.

Chamberlain J., Poesio M., Kruschwitz U., « A new life for a dead parrot : Incentive structures in the Phrase Detectives game », *Actes de WWW 2009*, Madrid, Espagne, avril, 2009.

Cohen K. B., Xia J., Zweigenbaum P., Callahan T., Hargraves O., Goss F., Ide N., Névéol A., Grouin C., Hunter L. E., « Three Dimensions of Reproducibility in Natural Language Processing », *Actes de 11th International Conference on Language Resources and Evaluation (LREC'18)*, Miyazaki, Japon, mai, 2018.

Colot S., Ludwig R., « Guadeloupean and Martinican Creole », *in* S. M. Michaelis, P. Maurer, M. Haspelmath, M. Huber (eds), *The survey of pidgin and creole languages.*, vol. 2, Oxford University Press, 2013.

Cox C., « Probabilistic tagging of minority language data : a case study using Qtag. », *Language & Computers*, vol. 71, n° 1, p. 213-231, 2010.

Crévenat-Werner D., Zeidler E., *Orthographe alsacienne - Bien écrire l'alsacien de Wissembourg à Ferrette*, Jérôme Do Bentzinger, 2008.

Dandapat S., Biswas P., Choudhury M., Bali K., « Complex Linguistic Annotation — No Easy Way out ! : A Case from Bangla and Hindi POS Labeling Tasks », *Actes de Linguistic Annotation Workshop*, ACL-IJCNLP '09, Stroudsburg, PA, États-Unis, p. 10-18, août, 2009.

Denis P., Sagot B., « Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morphosyntaxique état-de-l'art du français », *Actes de Traitement Automatique des Langues Naturelles (TALN'10)*, Montréal, Canada, juillet, 2010.

Diki-Kidiri M., « Comment assurer la présence d'une langue dans le cyberespace », *UNESCO. Retrieved December*, vol. 31, p. 2007, 2007.

Fišer D., Tavčar A., Erjavec T., « sloWCrowd : a Crowdsourcing Tool for Lexicographic Tasks », *Actes de 9th International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Islande, mai, 2014.

Fort K., Adda G., Cohen K. B., « Amazon Mechanical Turk : Gold Mine or Coal Mine ? », *Computational Linguistics (editorial)*, vol. 37, n° 2, p. 413-420, juin, 2011.

Fort K., Guillaume B., Lefèbvre N., « Who wants to play Zombie ? A survey of the players on ZOMBILINGO », *Actes de Games4NLP 2017 - Using Games and Gamification for Natural Language Processing*, Symposium Games4NLP, Valence, Espagne, p. 2, avril, 2017.

Fort K., Sagot B., « Influence of Pre-annotation on POS-tagged Corpus Development », *Actes de ACL Linguistic Annotation Workshop*, Uppsala, Suède, p. 56-63, juillet, 2010.

Garcia M., Gamallo P., Gayo I., Cruz M. A., « PoS-tagging the web in portuguese. National varieties, text typologies and spelling systems », *Procesamiento de Lenguaje Natural*, vol. 53, p. 95-101, 2014.

Garrette D., Mielens J., Baldridge J., « Real-World Semi-Supervised Learning of POS-Taggers for Low-Resource Languages », *Actes de 51st Annual Meeting of the Association for Computational Linguistics (ACL-2013)*, ACL '13, Sofia, Bulgarie, p. 583-592, août, 2013.

Guillaume B., Fort K., Lefebvre N., « Crowdsourcing Complex Language Resources : Playing to Annotate Dependency Syntax », *Actes de International Conference on Computational Linguistics (COLING)*, Osaka, Japon, décembre, 2016.

Hana J., Feldman A., Brew C., « A Resource-light Approach to Russian Morphology : Tagging Russian using Czech resources », *Actes de Conference on Empirical Methods in Natural Language Processing (EMNLP)*, ACL, Barcelone, Espagne, p. 222-229, juillet, 2004.

Hazaël-Massieux M.-C., *Ecrire en créole : Oralité et écriture aux Antilles*, L'Harmattan, 2000.

Hovy D., Plank B., Søgaard A., « Experiments with crowdsourced re-annotation of a POS tagging data set », *Actes de 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, Baltimore, MD, États-Unis, p. 377-382, juin, 2014.

Jamatia A., Das A., « Part-of-Speech Tagging System for Indian Social Media Text on Twitter », *Actes de Workshop on Language Technologies For Indian Social Media (SOCIAL-INDIA)*, Goa, Inde, p. 21-28, novembre, 2014.

Klubička F., Ljubešić N., « Using crowdsourcing in building a morphosyntactically annotated and lemmatized silver standard corpus of Croatian », *Actes de 9th Language Technologies Conference*, Ljubljana, Slovénie, octobre, 2014.

Krumm J., Davies N., Narayanaswami C., « User-Generated Content », *IEEE Pervasive Computing*, vol. 7, n° 4, p. 10-11, octobre, 2008.

Lafourcade M., Joubert A., « JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes », *Actes de Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, Lyon, France, mars, 2008.

Leemann A., Kolly M.-J., Goldman J.-P., Dellwo V., Hove I., Almajai I., Grimm S., Robert S., Wanitsch D., « Voice Äpp : a mobile app for crowdsourcing Swiss German dialect data », *Actes de INTERSPEECH 2015*, Dresde, Allemagne, septembre, 2015.

Li S., Graça J. a. V., Taskar B., « Wiki-ly Supervised Part-of-speech Tagging », *Actes de 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju, Corée du Sud, p. 1389-1398, juillet, 2012.

Ljubešic N., Zupan K., Fišer D., Erjavec T., « Normalising Slovene data : historical texts vs. user-generated content », *Actes de 13th Conference on Natural Language Processing (KONVENS 2016)*, Bochum, Allemagne, p. 146-155, septembre, 2016.

Ludwig R., Montbrand D., Poullet H., Telchid S., « Abrégé de grammaire du créole guadeloupéen », *Dictionnaire créole français (Guadeloupe), avec un abrégé de grammaire créole et un lexique français-créole*, SERVEDIT, p. 17-38, 1990.

Malherbe M., *Les langages de l'humanité (une encyclopédie des 3000 langues parlées dans le monde)*, Collection Bouquins, Laffont, 1983.

McEnery T., Hardie A., *Corpus Linguistics : Method, Theory and Practice*, Cambridge Textbooks in Linguistics, Cambridge University Press, 2011.

Melero M., Costa-Jussà M. R., Domingo J., Marquina M., Quixal M., « Holaaa ! ! writin like u talk is kewl but kinda hard 4 NLP », *Actes de 8th International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turquie, mai, 2012.

Millour A., Fort K., « Why do we Need Games ? Analysis of the Participation on a Crowdsourcing Annotation Platform », *Actes de Games4NLP 2017 - Using Games and Gamification for Natural Language Processing*, Symposium Games4NLP, Valence, Espagne, avril, 2017.

Millour A., Fort K., « Krik : First Steps into Crowdsourcing POS tags for Kréyòl Gwadloupéyen », *Actes de 11th International Conference on Language Resources and Evaluation (LREC'18)*, Workshop CCURL, Miyazaki, Japon, mai, 2018a.

Millour A., Fort K., « Toward a Lightweight Solution for Less-resourced Languages : Creating a POS Tagger for Alsatian Using Voluntary Crowdsourcing », *Actes de 11th International Conference on Language Resources and Evaluation (LREC'18)*, Miyazaki, Japon, mai, 2018b.

Munro R., « Crowdsourcing and the Crisis-Affected Community : lessons learned and looking forward from Mission 4636 », *Journal of Information Retrieval*, 2013.

Petrov S., Das D., McDonald R., « A Universal Part-of-Speech Tagset », *Actes de 8th International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turquie, mai, 2012.

Pingali S., Mortensen D., Littell P., Levin L., Phonetically-Aware Approximate Search for Low-Resource Languages, Technical report, Carnegie Mellon University, Pittsburgh, PA, États-Unis, 2017.

Plank B., « What to do about non-standard (or non-canonical) language in NLP », *Actes de 13th Conference on Natural Language Processing (KONVENS)*, Bochum, Allemagne, p. 13-20, août, 2016.

Prado D., « Présence des langues dans le monde réel et le cyberespace », *Net.lang Réussir le cyberespace multilingue*, c&f édition edn, Vannini, Laurent and Le Crosnier, Hervé, 2012.

Rivron V., « L'usage de Facebook chez les Éton du Cameroun », *Net.lang Réussir le cyberespace multilingue*, c&f édition edn, Vannini, Laurent and Le Crosnier, Hervé, p. 171-178, 2012.

Sagot B., Informatiser le lexique, Habilitation à diriger des recherches en inguistique informatique, Institut national de recherche en informatique et en automatique (Inria), juin, 2018.

Samardzic T., Scherrer Y., Glaser E., « Normalising orthographic and dialectal variants for the automatic processing of Swiss German », *Actes de 7th Language and Technology Conference*, Poznań, Pologne, novembre, 2015.

Scherrer Y., Sagot B., « Lexicon induction and part-of-speech tagging of non-resourced languages without any bilingual resources », *Actes de Workshop on Adaptation of language resources and tools for closely related languages and language variants*, RANLP '13, Hissar, Bulgarie, septembre, 2013.

Steiblé L., Bernhard D., « Pronunciation Dictionaries for the Alsatian Dialects to Analyze Spelling and Phonetic Variation », *Actes de 11th edition of the Language Resources and Evaluation Conference*, Miyazaki, Japon, mai, 2018.

Toutanova K., Klein D., Manning C. D., Singer Y., « Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network », *Actes de Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Stroudsburg, PA, États-Unis, p. 173-180, mai, 2003.

Tseng H., Jurafsky D., Manning C., « Morphological features help POS tagging of unknown words across language varieties », *Actes de Fourth SIGHAN Workshop on Chinese Language Processing*, Jeju, Corée du Sud, p. 32-39, octobre, 2005.

Tuite K., « GWAPs : Games with a Problem », *Actes de 9th International Conference on the Foundations of Digital Games*, Liberty of the Seas, Caraïbes, avril, 2014.

Zaghouani W., Dukes K., « Can Crowdsourcing be used for Effective Annotation of Arabic ? », *Actes de 9th International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Islande, mai, 2014.

Analyse syntaxique de langues faiblement dotées à partir de plongements de mots multilingues

Application au same du nord et au komi-zyriène

KyungTae Lim — Niko Partanen — Thierry Poibeau

Laboratoire LATTICE
CNRS et École normale supérieure, PSL et Université Sorbonne nouvelle, USPC
1 rue Maurice Arnoux, 92120 Montrouge – France
prenom.nom@ens.fr

RÉSUMÉ. *Cet article présente une tentative pour appliquer des méthodes d'analyse syntaxique performantes, à base de réseaux de neurones récursifs, à des langues pour lesquelles on dispose de très peu de ressources. Nous proposons une méthode originale à base de plongements de mots multilingues obtenus à partir de langues plus ou moins proches typologiquement, afin de déterminer la meilleure combinaison de langues possibles pour l'apprentissage. L'approche a permis d'obtenir des résultats encourageants dans des contextes considérés comme linguistiquement difficiles. Le code source est disponible en ligne (voir https://github.com/jujbob).*

ABSTRACT. *This article presents an attempt to apply efficient parsing methods based on recursive neural networks to languages for which very few resources are available. We propose an original approach based on multilingual word embeddings acquired from different languages so as to determine the best language combination for learning. The approach yields competitive results in contexts considered as linguistically difficult.*

MOTS-CLÉS : *analyse syntaxique, modèles multilingues, plongements de mots, langues peu dotées, same du nord, komi-zyriène*

KEYWORDS: *parsing, multilingual models, word embeddings, low-resource languages, North Saami, Komi-Zyrian*

1. Introduction

Le développement de systèmes automatiques, pouvant analyser avec succès des langues faiblement dotées, est une question cruciale pour le traitement automatique des langues (TAL). La plupart des systèmes d'analyse sont en effet fondés sur des techniques d'apprentissage supervisé nécessitant de grandes quantités de données annotées : la disponibilité de tels corpus est une des conditions principales pour obtenir des performances correctes, quelle que soit la tâche visée. Ce type de techniques est donc bien adapté pour les quelques langues pour lesquelles on dispose de nombreuses ressources en ligne (dictionnaires et surtout corpus annotés), mais l'approche laisse aussi de nombreuses autres langues de côté, du fait de l'absence des ressources nécessaires. Par ailleurs, produire des données annotées en grandes quantités demande beaucoup de moyens (que ce soit au niveau humain ou financier). C'est évidemment un problème majeur pour quantité de langues pour lesquelles ces données sont quasi inexistantes et pour lesquelles on ne dispose pas des moyens nécessaires pour y remédier.

Nous nous intéressons dans cet article au cas de l'analyse syntaxique (cet article reprend en partie la présentation du système développé par le LATTICE pour la tâche d'évaluation CoNLL 2017 (Lim et Poibeau, 2017 ; Lim *et al.*, 2018))[1]. L'analyse syntaxique est une tâche classique, fondamentale pour le TAL et nécessaire pour de nombreuses applications dérivées. Les systèmes d'analyse syntaxique récents les plus performants ou les plus emblématiques du domaine (Weiss *et al.*, 2015 ; Straka *et al.*, 2016 ; Ballesteros *et al.*, 2016), pour n'en citer que quelques-uns, reposent tous sur l'approche décrite dans le paragraphe précédent, c'est-à-dire sur une approche par apprentissage supervisé à partir de grands corpus annotés de la langue visée, souvent l'anglais.

La communauté a toutefois bien conscience qu'il faut aller au-delà des quelques langues bien dotées pour lesquelles on dispose de ressources en masse. D'une part parce qu'il y a des besoins concrets pour d'autres langues : différentes communautés linguistiques, en particulier celles liées à des langues minoritaires ou en danger, ont conscience que l'avenir passe entre autres par l'informatisation des langues et la mise au point d'outils performants, y compris pour le grand public. D'autre part, parce que les langues moins bien dotées posent souvent des questions extrêmement intéressantes sur le plan linguistique, et qui ont été trop longtemps négligées jusqu'ici. Les systèmes entraînés seulement sur l'anglais ne donnent qu'une vision étriquée du TAL, visant l'analyse d'une langue analytique à la morphologie extrêmement pauvre. Au-delà de la quantité de données disponible, la prise en compte de la complexité linguistique, notamment morphosyntaxique, est un autre élément fondamental.

Pour prendre un exemple récent, la tâche d'évaluation lors des conférences *Computational Natural Language Learning* 2017 et 2018 (*CoNLL shared task*) (Zeman,

1. Le code source correspondant aux réalisations présentées dans cet article est intégralement disponible sur le site : `https://github.com/jujbob`.

D. *et al.*, 2017 ; Zeman *et al.*, 2018) portait sur environ cinquante langues (plus précisément quarante-neuf en 2017 et cinquante-sept en 2018), soit à peu près toutes les langues pour lesquelles des données annotées syntaxiquement sont disponibles en quantité significative au format UD (Universal Dependencies) (Nivre *et al.*, 2016). C'est probablement le défi d'analyse syntaxique le plus ambitieux jamais organisé de ce point de vue, mais le chiffre de cinquante langues est à considérer en regard des six mille langues estimées dans le monde. Et même si l'on ne considère que les langues pour lesquelles des données écrites sont disponibles, les cinquante-sept langues de CoNLL 2018 ne permettent de couvrir qu'un échantillon extrêmement restreint de ce qui existe.

En ce qui concerne l'analyse syntaxique automatique, l'approche monolingue et supervisée (c'est-à-dire par apprentissage automatique à partir de corpus annotés représentatifs) est évidemment la plus répandue (l'alternative étant les systèmes reposant sur des grammaires entièrement élaborées à la main). Des chercheurs essaient toutefois depuis un certain temps de concevoir des systèmes multilingues. L'approche multilingue a donné des résultats encourageants aussi bien pour les langues faiblement dotées (Guo *et al.*, 2015 ; Guo *et al.*, 2016) que pour les langues déjà bien dotées, et disposant déjà de ressources comme des dictionnaires et des corpus représentatifs (Ammar *et al.*, 2016a ; Ammar *et al.*, 2016b). Dans ce dernier cas, l'idée est de n'avoir à maintenir qu'un seul modèle d'analyse et de pouvoir l'appliquer ensuite aux différentes langues constitutives du modèle. L'approche multilingue a de plus un avantage, même pour les langues bien dotées : en mettant ensemble plusieurs langues, on peut espérer mieux analyser certains mots ou certaines constructions rares sans dégrader les performances sur des phénomènes plus classiques. Il semble donc y avoir toujours un gain possible.

Ainsi, Ammar et ses collègues (cf. références déjà citées) ont proposé des études portant sur des langues indo-européennes pour lesquelles on dispose déjà de ressources importantes. Ils ont démontré que l'approche *via* un modèle multilingue donne généralement de meilleurs résultats que les modèles monolingues correspondants pour les langues visées.

D'une manière générale, c'est surtout pour les langues moins bien dotées que l'approche multilingue est intéressante. Cette approche peut en fait être mise en œuvre de deux façons différentes. La première consiste à projeter des annotations disponibles d'une langue donnée vers une langue peu dotée *via* un corpus parallèle. Cette approche a été utilisée à plusieurs reprises (notamment dans les références déjà citées (Guo *et al.*, 2015 ; Ammar *et al.*, 2016b)), mais elle nécessite de disposer de données parallèles en quantité suffisante, ce qui est souvent problématique. Le transfert de connaissances nous semble aussi problématique en soi, dans la mesure où cela suppose une relative similarité de structure entre les deux langues visées. Si les deux langues sont trop différentes, l'approche fonctionnera mal, ce qui est insatisfaisant à la fois sur le plan pratique et sur le plan théorique. La seconde approche, celle que nous adoptons ici, vise à produire directement un modèle multilingue, pouvant fonctionner pour plusieurs langues, tout en relâchant les contraintes de structure (voir aussi

Scherrer et Sagot (2014) pour une expérience visant l'étiquetage morphosyntaxique de langues non dotées par transfert depuis une langue dotée, sans utilisation de corpus parallèles).

Dans cet article, nous proposons une approche d'analyse syntaxique utilisant des méthodes à l'état de l'art pour des langues disposant de très peu de ressources structurées (mais pour lesquelles des corpus bruts, c'est-à-dire non annotés, sont disponibles). Notre approche ne nécessite qu'un petit dictionnaire bilingue (ou, *a minima*, une liste élaborée manuellement de mots de la langue visée avec leur traduction dans la langue cible) et l'annotation syntaxique (au format UD) manuelle d'une poignée de phrases de la langue visée. Ces données peuvent être mises au point en quelques heures seulement (moins d'une journée) par une personne connaissant la langue en question. Comme souvent dans ce type de schéma, l'hypothèse que nous faisons est qu'il est possible de transférer des connaissances d'une langue à l'autre entre langues apparentées, mais nous ne faisons pas pour autant l'hypothèse d'une similarité de structure stricte entre les langues. Le point principal est d'identifier des éléments communs au niveau lexical, *via* des plongements de mots (*word embeddings*) multilingues. La source première de comparaison entre une phrase en langue source et une phrase en langue cible est donc lexicale et sémantique, plus que syntaxique (même si la syntaxe joue aussi un rôle primordial, bien évidemment ; c'est d'ailleurs pour cela que les langues choisies pour élaborer le modèle d'analyse doivent être sélectionnées avec attention).

Nous faisons aussi l'hypothèse que les performances dépendent largement des langues utilisées pour élaborer le modèle d'analyse. *A priori*, des langues de même famille et, au sein d'une même famille, des langues étroitement apparentées sont évidemment les meilleurs candidats *a priori*, mais les contacts linguistiques peuvent aussi jouer un rôle. Il existe en effet de nombreux cas de langues où les locuteurs sont tous au moins bilingues et s'expriment le plus souvent dans la langue « dominante », ce qui peut affecter largement leur langue maternelle. Ces phénomènes sont connus, mais relativement peu étudiés, et le TAL peut aider à donner une base statistique et quantitative à l'étude de ces phénomènes d'emprunts et de contagion linguistique. Au cours de l'étude, nous détaillerons plusieurs modèles mettant en jeu des langues génétiquement apparentées et non apparentées, afin de mieux comprendre les limites ou les possibilités de transfert de modèles entre différentes familles de langues.

Afin de mener à bien nos expériences, nous nous penchons sur deux langues finno-ougriennes peu dotées et ayant fait l'objet de peu de recherches en TAL jusqu'ici. Le same du nord[2] est une langue parlée par environ 20 000 personnes au nord de la péninsule scandinave (Suède, Norvège, Finlande). Il existe une dizaine de langues sames (25 000 à 30 000 locuteurs environ au total), mais le same du nord est de loin la langue la plus répandue et celle qui est la mieux supportée par les autorités et les médias (il existe des journaux, ainsi qu'une radio et des émissions de télévision soutenues pu-

2. glottolog.org/resource/languoid/id/nort2671

bliquement). Nous nous intéressons par ailleurs au komi-zyriène[3] (parfois abrégé en komi par la suite), une langue finno-ougrienne de Russie assez éloignée du same. Quasiment tous les locuteurs komis parlent aussi le russe, qui est leur langue essentielle de communication (souvent même entre locuteurs komis). Il y a environ 150 000 locuteurs komis.

Les deux langues (same du nord et komi) sont dans des situations différentes mais possèdent aussi de nombreux points communs pour leur avenir : tous les locuteurs sont bilingues, ils utilisent essentiellement une autre langue de communication (le russe pour les Komis, le finnois, le suédois ou le norvégien pour les Sames du nord) et le komi comme le same du nord étaient dévalorisés jusqu'à récemment. La situation a toutefois changé depuis les années 1980 : les communautés ont pris conscience de l'importance de la préservation de leur langue maternelle, des campagnes de numérisation ont permis de rendre disponibles les écrits existants (la production écrite disponible s'étend sur un siècle environ) et surtout la langue est transmise activement aux enfants, au moins dans certaines régions et dans certaines communautés. Les Sames comme les Komis sont aujourd'hui convaincus de l'importance de développer des outils informatiques pour aider à maintenir et développer leur langue.

Sur le plan informatique, la situation des deux langues n'est pas la même. Le centre d'analyse linguistique de l'université de Tromsø (projet Giellatekno[4]) développe depuis plusieurs années des outils permettant la description des langues finno-ougriennes en général, et du same en particulier. On dispose donc de dictionnaires électroniques assez complets pour le same du nord, incluant les paradigmes de flexion et de conjugaison, ce qui permet d'un côté de générer l'essentiel des formes de la langue et de l'autre, d'analyser dynamiquement des formes linguistiques complexes. Les outils d'analyse (analyse morphosyntaxique et syntaxique) sont en revanche limités, l'approche de l'équipe de Tromsø reposant uniquement sur des automates à nombre fini d'états (éventuellement pondérés), mais sans recours à l'apprentissage automatique. Pour le komi, la situation est beaucoup moins favorable que pour le same du nord. L'équipe de Tromsø a commencé à décrire le komi mais les données disponibles restent relativement embryonnaires.

Il faut enfin noter que, fin 2017, un corpus annoté au format UD a été rendu disponible pour le same du nord. Les données disponibles pour le same sont donc aujourd'hui suffisamment massives pour pouvoir évaluer précisément des analyseurs pour cette langue, mais aussi pour développer des analyseurs de manière traditionnelle, à partir d'un corpus d'entraînement important, comme lors de la campagne CoNLL 2018. Pour le komi, la situation est très différente et il n'existait, à notre connaissance, aucun corpus syntaxiquement annoté pour cette langue au moment où nous avons commencé nos expériences. Les quelques corpus électroniques disponibles sont liés à des travaux réalisés à des fins de documentation linguistique (Blokland *et al.*, 2015 ; Gerstenberger *et al.*, 2016) : ces corpus sont relativement petits et dif-

3. `glottolog.org/resource/languoid/id/komi1268`
4. `http://giellatekno.uit.no/`

ficiles d'utilisation dans une perspective de TAL. Notons enfin que ces langues disposent de données numérisées en quantités relativement importantes, ce qui est utile pour l'élaboration de plongements de mots et permet par ailleurs de compenser en partie le manque de ressources.

L'article est structuré comme suit : nous présentons dans un premier temps l'état de l'art en matière d'analyse syntaxique multilingue (section 2). Nous présentons ensuite le modèle lexical mis au point pour nos expériences (section 3), puis l'architecture du modèle d'analyse à base de réseaux de neurones bidirectionnels, dit BiLSTM (section 4). Nous présentons ensuite le détail des expériences sur le same du nord et le komi-zyriène (section 5), avant de finir par une discussion de ces résultats (section 6), une conclusion et quelques perspectives (section 7). Nous présentons enfin les données mises au point pour le komi-zyriène (embryon de corpus annotés au format Universal Dependencies), ainsi que quelques exemples en annexe à cet article.

2. État de l'art

Depuis les travaux pionniers de Hwa *et al.* (2005), de nombreux groupes se sont intéressés à la mise au point d'analyseurs syntaxiques multilingues, et/ou au transfert de connaissances d'une langue à l'autre, que ce soit dans un cadre d'analyse syntaxique ou pour d'autres tâches, par exemple l'analyse morphosyntaxique. La plupart des méthodes supposent un corpus parallèle, avec des annotations d'un côté (langue source), et non de l'autre (langue cible). La tâche repose alors le plus souvent sur une stratégie de transfert d'étiquettes (c'est-à-dire d'annotations) d'une langue à l'autre, en tenant compte des spécificités de chaque langue. D'autres approches évitent le transfert direct en proposant des stratégies plus ou moins élaborées visant tout d'abord à produire des représentations multilingues avancées, pour éviter les problèmes de transfert d'information. L'apprentissage du parseur est alors réalisé directement sur le modèle enrichi ainsi défini.

Comme on l'a dit, les approches reposant sur la projection d'annotations utilisent un corpus parallèle annoté dans la langue source. Ces annotations sont projetées sur le corpus en langue cible, à partir de quoi un analyseur syntaxique peut être inféré par apprentissage automatique (Smith et Eisner, 2009 ; Zhao *et al.*, 2009 ; Liu *et al.*, 2013). Cette approche est efficace mais elle est principalement confrontée à des problèmes liés à l'alignement des mots lors de l'étape de projection d'annotations. Les méthodes proposées reposent sur des algorithmes de projection robustes prenant en compte un contexte large (Das et Petrov, 2011), ou sur des ressources extérieures comme Wikipédia (Kim *et al.*, 2014) ou WordNet (Khapra *et al.*, 2010), ou bien encore sur la correction *a posteriori* de certaines étiquettes de manière heuristique (Kim *et al.*, 2010).

L'alternative consiste à élaborer directement des modèles d'analyse multilingues grâce aux informations contenues dans des corpus parallèles, ou grâce à des connaissances extérieures, provenant en général de dictionnaires bilingues. L'approche consiste à « apprendre » un modèle d'analyse unique, conjointement pour les deux

langues. Des règles, spécifiées ou non à la main, permettent ensuite d'adapter l'analyse et de tenir compte des spécificités des langues considérées. En dehors de l'analyse syntaxique, les modèles multilingues ont été appliqués à d'autres problèmes de traitement automatique des langues, comme la reconnaissance des entités (Zhuang et Zong, 2010) ou l'analyse des rôles sémantique (Kozhevnikov et Titov, 2012).

D'autres méthodes enfin empruntent aux deux approches précédentes pour créer un modèle d'analyse hybride. Il s'agit alors de produire dans un premier temps une représentation en grande partie indépendante des langues (ou plutôt mêlant les différentes langues dans un seul espace de représentation partagé) puis à « apprendre » un analyseur à partir de cette représentation abstraite et « *crosslingue* » (Täckström *et al.*, 2012). Différents types de ressources peuvent être utilisés dans ce cadre, notamment des corpus parallèles et/ou des dictionnaires bilingues.

Les systèmes plus récents reposent quasi systématiquement sur la notion de plongement de mots (« *word embeddings* » en anglais). Comme précédemment, les systèmes utilisent soit des dictionnaires soit des corpus bilingues, voire des documents parallèles (des légendes d'images ou des pages Wikipédia, par exemple) comme source de connaissances pour inférer un modèle bilingue. Une grande variété d'approches a pu être proposée, mais plusieurs auteurs ont montré que ce sont les données utilisées pour l'apprentissage, plus que l'architecture ou les algorithmes utilisés, qui ont une influence majeure sur le résultat final (Levy *et al.*, 2017 ; Ruder *et al.*, 2017). En gros, à partir des mêmes données, on obtient des résultats très similaires avec des approches en apparence différentes, car dans les faits les algorithmes eux-mêmes sont au final relativement similaires, quel que soit leur point de départ.

L'article de Ruder *et al.* (2017) présente en détail les méthodes fondées sur des représentations lexicales riches. Trois approches sont possibles pour obtenir des plongements de mots bilingues (ou multilingues si on généralise l'approche) : *i*) une première approche consiste à obtenir des représentations sous forme de plongements de mots indépendants pour les deux langues visées (selon la technique introduite par Mikolov *et al.* (2013a) par exemple), puis à mettre en relation les deux représentations obtenues par projection d'un espace sémantique sur l'autre, comme par exemple dans (Artetxe *et al.*, 2016) ; *ii*) élaborer directement un modèle bilingue à partir d'un corpus dans lequel des phrases (voire des documents) des deux langues visées sont déjà en rapport direct (corpus parallèle ou similaire) (Gouws et Søgaard, 2015 ; Gouws *et al.*, 2015) ou *iii*) utiliser un corpus parallèle et un espace sémantique pour chaque langue simultanément (Luong *et al.*, 2015), afin d'obtenir la représentation la plus adéquate en fonction des données fournies en entrée au système.

La mise au point de plongements de mots bilingues et multilingues est un secteur clé de la recherche en TAL à l'heure actuelle. Les tendances visent à réduire les contraintes sur les données en entrée pour obtenir des approches rapides, efficaces et surtout simples à mettre en œuvre. Ainsi, Artexte *et al.* (2017) montrent que quelques dizaines (une cinquantaine environ) de couples de mots bien choisis sont suffisants pour obtenir des plongements de mots bilingues de bonne qualité, au lieu des quelques milliers utilisés dans les expériences précédentes. Une équipe de Facebook a même ré-

cemment montré qu'on pouvait produire des plongements de mots bilingues sans données parallèles ni thésaurus bilingue (Conneau *et al.*, 2017). Cet article a eu un relatif retentissement, mais ses conclusions doivent être nuancées, les résultats n'étant satisfaisants que si les corpus utilisés sont très proches stylistiquement et thématiquement (Vulić *et al.*, 2018).

Dans cet article, nous utiliserons la première méthode qui est facile à mettre en œuvre et qui semble obtenir des résultats très satisfaisants malgré sa simplicité. En ce qui concerne l'architecture du système, nous nous inspirons de Guo *et al.* (2015). La principale différence est que Guo et ses collègues utilisent une approche délexicalisée pour leur analyse, tandis que, conformément au système de Ammar *et al.* (2016a), nous avons recours à des représentations multilingues riches pour l'analyse.

3. Mise au point d'un modèle lexical multilingue

Dans la mesure où les langues que nous souhaitons analyser sont finno-ougriennes, nous nous tournons naturellement vers le finnois pour obtenir des connaissances pertinentes pour l'analyse. Le same du nord a été en contact depuis plusieurs siècles avec le finnois et ce sont surtout deux langues étroitement liées sur le plan génétique (Aikio, 2012, p. 67–69). Le komi est plus éloigné du finnois, mais le finnois reste la langue la plus proche sur le plan linguistique pour laquelle on dispose de ressources importantes. Nous nous sommes également intéressés au russe, sachant que le komi est depuis longtemps en contact avec le russe, et que tous les locuteurs komis sont bilingues (ils parlent aussi russe). On peut donc s'attendre à ce que le russe ait influencé le komi et que ce soit une autre source de connaissances pertinente, les structures copiées du russe étant fréquentes en komi, surtout à l'oral (Leinonen, 2006, p. 241).

Enfin, des expériences avec un corpus anglais seront aussi effectuées : l'anglais n'a pas de lien génétique avec le komi ou le same, ce qui le rend intéressant comme « langue de contrôle » (c'est-à-dire pour comparer les performances obtenues par rapport à des langues de la même famille linguistique, par exemple). Il faut toutefois faire attention aux expériences avec l'anglais : la masse de données disponible pour cette langue permet souvent d'obtenir des résultats relativement corrects malgré tout, la quantité permettant de suppléer partiellement au manque de qualité (ou du moins à l'absence de similarité entre l'anglais et les langues visées lors de l'analyse). L'anglais peut aussi avoir une influence bénéfique en apportant des éléments d'information pertinents pour le niveau lexical-sémantique, ce qui est utile même pour une tâche d'analyse syntaxique.

3.1. *Préparation de ressources linguistiques*

Comme nous l'avons déjà dit, pour les expériences qui suivent nous avons recours aux lexiques bilingues disponibles sur le site Giellatekno[5]. Nous avons par ailleurs utilisé les plongements de mots FastText proposés par Facebook en mai 2017 pour le finnois et le russe (Bojanowski *et al.*, 2017). Il nous faut ensuite générer des plongements de mots similaires pour le same et le komi. Pour ce faire, nous avons en premier lieu recours au corpus Wikipédia, mais il s'agit d'un corpus relativement petit pour les langues visées. Nous le complétons alors avec des corpus disponibles dans le domaine public[6]. Nous produisons enfin les plongements de mots monolingues pour chacune des langues considérées à partir du module FastText de Facebook.

3.2. *Projection de plongements de mots pour obtenir une ressource multilingue*

Dans la section précédente, nous avons décrit comment nous avons obtenu des plongements de mots monolingues pour chaque langue considérée mais, logiquement, chacun de ces plongements a son propre espace vectoriel. Afin d'obtenir des plongements de mots bilingues (voire multilingues, en répétant l'opération plusieurs fois), c'est-à-dire des plongements de mots partageant un espace vectoriel unique, nous utilisons la méthode de transformation linéaire proposée par Artexte *et al.* (2016). Pour effectuer cette transformation, il est nécessaire d'avoir un petit lexique bilingue qui va permettre de définir des « points d'attache » entre les deux espaces vectoriels à mettre en regard. Selon les comparaisons présentées dans Artexte *et al.* (2017, p. 457), la taille des dictionnaires que nous utilisons ici est bien supérieure à ce qui est nécessaire pour effectuer la mise en correspondance des deux espaces vectoriels des langues concernées (tableau 1). Il serait intéressant d'essayer avec de très petits dictionnaires, de quelques dizaines de mots au maximum, afin d'estimer la dégradation des performances dans ce cas de figure, mais comme nous disposons de dictionnaires bilingues contenant plusieurs milliers de mots, nous n'avons pas à ce stade exploré de contextes plus difficiles (mais cela sera nécessaire si l'on doit s'intéresser à d'autres langues ouraliennes, moins bien dotées que le same du nord ou le komi-zyriène).

La méthode de projection des deux espaces sémantiques l'un sur l'autre est la suivante. Soit deux plongements de mots différents, l'un X correspondant à la langue

5. Les dictionnaires pour le same sont disponibles ici : `http://dicts.uit.no/smedicts.eng.html` et les autres dictionnaires sont disponibles à l'adresse suivante : `https://gtsvn.uit.no/langtech/trunk/words/dicts/`. Tous ces dictionnaires sont relativement complets et disponibles sous forme libre, avec une licence GNU GPLv3.

6. Notamment les livres numérisés de la collection Fenno-Ugrica (`https://fennougrica.kansalliskirjasto.fi/`) qui ont été corrigés manuellement par le laboratoire d'appui à la production de ressources électroniques pour les langues régionales de Syktyvkar (`http://komikyv.org/`). Pour le same du nord, nous utilisons le corpus gratuit SIKOR (`http://hdl.handle.net/11509/100`), disponible avec une licence CC-BY 3.0.

Paire linguistique	**Taille du dictionnaire bilingue (nombre de couples de mots)**	**Taille des plongements de mots correspondants**
Finnois-same du nord	12 398	2,4 Go
Same du nord-finnois	10 541	2,4 Go
Same du nord-anglais	1 499	1,4 Go
Finnois-komi	12 879	2,3 Go
Komi-anglais	8 746	7,5 Go
Russe-komi	12 354	5,7 Go

Tableau 1. *Taille des dictionnaires et des plongements de mots liés générés à partir des différents dictionnaires (il s'agit de dictionnaires de formes fléchies, ce qui explique que la taille du dictionnaire finnois-same du nord soit par exemple différente de celle du dictionnaire same du nord-finnois).*

cible, et l'autre Y à la langue source, et soit $D=\{(x_i,y_i)\}_{i=1}^{m}$ (*où* $x_i \in X$, $y_i \in Y$) la ressource obtenue consistant en une collection de plongements de mots bilingues. Le but est, dès lors, de trouver la matrice de transformation W telle que xW soit la meilleure approximation de y. On obtient ce résultat en minimisant la somme des carrés des erreurs, suivant Mikolov *et al.* (2013b) :

$$\arg\min_{W} \sum_{i=1}^{m} \|x_i W - y_i\|^2 \qquad [1]$$

Une dégradation importante des résultats peut se produire si la transformation linéaire est appliquée à deux plongements de mots sans autre contrainte. Pour répondre à ce problème, Artetxe *et al.* (2016) proposent une méthode de correspondance orthogonale qui permet de garder un niveau de performance correct. C'est cette variante de l'algorithme que nous avons utilisée ici.

3.3. *Corpus annotés au format Universal Dependencies*

Nous avons également besoin de corpus annotés pour nos expériences, au moins pour montrer leur apport quand ils sont disponibles. Nous avons utilisé des corpus pour l'anglais, le finnois et le russe : tous provenaient de l'initiative Universal Dependencies et peuvent être trouvés en ligne[7].

7. Sur le projet Universal dependencies, voir `http://universaldependencies.org`. Nous avons utilisé les corpus arobés suivants, dans leur version 2.1 : `https://github.com/UniversalDependencies/UD_English-EWT` (anglais), `https://github.com/UniversalDependencies/UD_Russian-GSD` (russe) et `https://github.com/UniversalDependencies/UD_Finnish-TDT` (finnois).

4. Modèle d'analyse en dépendances *crosslingue*

Les analyseurs syntaxiques traditionnels emploient des méthodes d'apprentissage supervisé fondées sur des séries de traits définis en grande partie manuellement. Le développeur doit en fait définir des combinaisons pertinentes (*feature functions*) de traits et de relations entre ceux-ci, afin que le système soit capable de déterminer les relations entre têtes et dépendants[8]. La définition manuelle de ces combinaisons de traits est une tâche difficile et en grande partie arbitraire, que tous les concepteurs de systèmes cherchent à contourner.

Les systèmes récents à base de réseaux de neurones ont plutôt recours à des méthodes automatiques permettant de simplifier le problème, en laissant le soin à la machine de déterminer les combinaisons de traits pertinentes. Ainsi, Chen et Manning (2014) ont proposé d'utiliser des classificateurs non linéaires intégrés dans un modèle de réseau neuronal. Avec cette méthode, les caractéristiques lexicales et non lexicales sont encodées dans des vecteurs qui peuvent être concaténés pour alimenter un classificateur non linéaire. Cette approche présente deux avantages : *i*) les classificateurs non linéaires ont globalement de meilleures performances que les classifieurs linéaires pour identifier les relations entre les éléments pertinents pour l'analyse, et *ii*) cette approche réduit drastiquement le travail manuel dans la mesure où le réseau de neurones se fonde essentiellement sur les caractéristiques calculées par les classifieurs.

4.1. *Architecture du système d'analyse*

Notre approche est ici similaire à celle de Chen et Manning (2014) et de Kiperwasser et Goldberg (2016) pour la partie analyse, mais nous utilisons des plongements de mots multilingues, alors que nos prédécesseurs s'en tiennent à un système monolingue.

Représentations LSTM bidirectionnelles. Les progrès récents en TAL sont largement dus à des représentations sous forme de traits portant des informations efficaces pour l'analyse des relations entre les mots de la phrase (Cho, 2015 ; Huang *et al.*, 2015). Une représentation LSTM bidirectionnelle (bi-LSTM) est un type de réseau de neurones récurrent, où chaque élément dans la séquence à analyser est lui-même représenté par un vecteur. L'algorithme procède en produisant des représentations préfixes (dites *forward* car la phrase est analysée de gauche à droite) et des représentations suffixes (dites *backward* car la phrase est alors analysée de droite à gauche). Un item est représenté par la concaténation de ses deux contextes, gauche

8. "*Traditionally, state-of-the-art parsers rely on linear models over hand-crafted feature functions. The feature functions look at core components (e.g. "word on top of stack", "leftmost child of the second-to- top word on the stack", "distance between the head and the modifier words"), and are comprised of several templates, where each template instantiates a binary indicator function over a conjunction of core elements (resulting in features of the form "word on top of stack is X and leftmost child is Y and ...*")." (Kiperwasser et Goldberg, 2016).

et droit. Soit par exemple la phrase $t = (t_1,t_2,...,t_n)$, dans laquelle le symbole $\circ$ dénote une opération de concaténation. La fonction LSTM bidirectionnelle correspond à : $\text{BiLSTM}(t_{1:n}, i) = \text{LSTM}_{Forward}(t_{1:i}) \circ \text{LSTM}_{Backward}(t_{i:n})$.

L'architecture est exactement la même que celle du BIST-parser (Kiperwasser et Goldberg, 2016). Nous renvoyons donc le lecteur à cet article fondateur pour connaître les détails de l'architecture du système qui est de ce point de vue relativement standard. Nous avons juste étendu cet analyseur de manière à le rendre multilingue, ce qui oblige à prendre en compte des représentations contextuelles construites par le module bi-LSTM multilingue (un code pour chaque mot, dit *one hot encoding*, permet de déterminer la langue associée).

Représentation lexicale. Soit une phrase en entrée $t = (t_1,t_2,...,t_n)$, une forme lexicale w, une étiquette morphosyntaxique correspondante p, un plongement de mots obtenu préalablement xw et une valeur de codage de la langue concernée l, un mot t_i (*token*) est défini comme : $t_i = e(w_i) \circ e(p_i) \circ e(xw_i) \circ e(l_i)$, où e réfère au plongement de chaque trait et $e(xw_i)$ est le plongement de mots déjà présenté en section 3. Nous ajoutons un code pour désigner la langue concernée, comme dit précédemment (Naseem *et al.*, 2012 ; Ammar *et al.*, 2016a). La plupart des analyseurs monolingues utilisent des traits comme $e(w_i)$ et $e(p_i)$, ainsi que d'autres éléments comme la distance entre la tête et le dépendant, ou d'autres traits spécifiques calculables à partir du corpus UD. Notez enfin que t_i de $\text{BiLSTM}(t_{1:n}, i)$ permet de stocker les contextes *forward* et *backward* du LSTM.

4.2. *Modèle d'analyse*

Il existe deux approches principales en matière d'analyse syntaxique en dépendance. La première est fondée sur la notion de transition (Nivre, 2004), l'autre sur la notion de graphe (McDonald *et al.*, 2005b). Nous utilisons ici une approche à base de graphes héritée du BIST-parser. L'approche semble efficace pour les corpus au format Universal Dependencies, et on renverra à Dozat *et al.* (2017) pour une comparaison détaillée et argumentée des deux approches.

À partir des représentations des mots et de leurs annotations dans la couche BiLSTM, le BIST-parser produit un arbre candidat pour chaque couple dépendant-mot-tête. Les scores attachés aux différents arbres candidats sont ensuite calculés à l'aide d'un perceptron multicouche (MLP), utilisé comme simple fonction de pondération (*scoring function*). Enfin, le système choisit les meilleurs arbres d'analyse en dépendance sur la base de la somme des scores attachés aux différents sous-arbres. Pour plus d'informations sur le modèle à base de graphes et sur le modèle de pondération d'arcs utilisé par le BIST-parser, voir (Taskar *et al.*, 2005) et (McDonald *et al.*, 2005a).

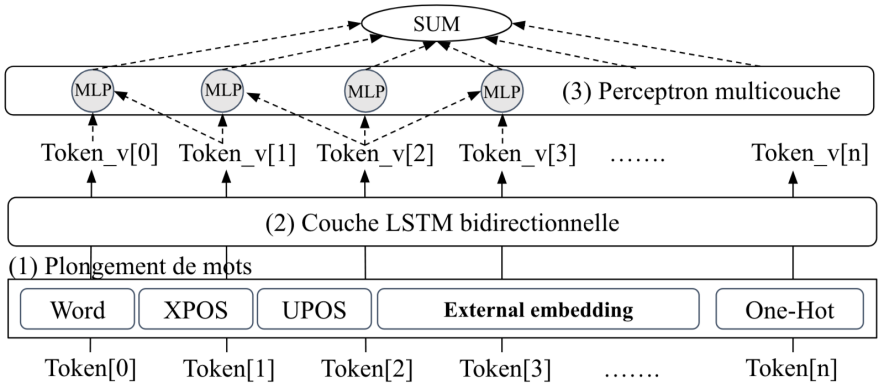


Figure 1. *Architecture du réseau de neurones*

5. Expériences

Nous présentons dans cette section les expériences que nous avons menées sur le same du nord et sur le komi-zyriène.

Corpus disponibles. Le same était une des langues dites *surprise language* de la campagne d'évaluation CoNLL 2017 : seulement vingt phrases étaient fournies pour l'entraînement et les participants n'avaient que quelques jours pour produire un système opérationnel. Le komi n'était pas inclus dans l'évaluation CoNLL 2017 mais nous avons choisi cette langue pour des raisons linguistiques (il s'agit d'une langue finno-ougrienne, comme le same) et parce qu'elle correspond à un cas typique de langue sous-dotée, comme on l'a vu dans l'introduction. Pour pouvoir mener à bien nos expériences, nous avons produit un corpus annoté composé de dix phrases komies pour l'entraînement et soixante-quinze phrases pour le test (ce corpus contient aujourd'hui près de trois cents phrases et grossit régulièrement, mais moins d'une centaine de phrases étaient disponibles au moment des expériences rapportées ici). Une présentation du corpus komi et des problèmes d'annotation rencontrés est disponible à la fin de cet article, en annexe.

Étiquettes morphosyntaxiques utilisées. Dans les expériences rapportées ici, nous nous fondons sur les étiquettes (catégories morphosyntaxiques) fournies par UDpipe (Straka *et al.*, 2016) pour le same et, en l'absence d'analyseur morphosyntaxique disponible pour le komi, nous utilisons les étiquettes posées à la main pour cette langue (*gold*). Lors de la campagne CoNLL 2017, l'analyse se fondait aussi sur des étiquettes de référence pour le same du nord, mais nous préférons recourir ici à un analyseur morphologique pour le same afin de rendre les conditions expérimentales plus proches de la réalité. Nous n'utilisons pas de traits morphologiques autres que ceux du corpus de référence ou ceux fournis par UDpipe pour le same.

Équipe	Score LAS	Score UAS
C2L2 (Ithaca)	48,96	58,85
IMS (Stuttgart)	40,67	51,56
HIT-SCIR (Harbin)	38,91	52,51
Notre système	**28,39**	**42,72**

Tableau 2. *Meilleurs résultats (officiels) pour le same lors de la tâche commune CoNLL 2017 et résultat obtenu par le LATTICE lors de cette même évaluation*

Le fait d'utiliser des étiquettes morphosyntaxiques de référence est bien évidemment quelque peu artificiel, mais permet de se focaliser uniquement sur le niveau syntaxique. Il n'en reste pas moins que la production d'analyseurs morphosyntaxiques performants est évidemment une condition nécessaire pour produire des analyses en situation réelle. La tâche commune CoNLL 2018 impliquait de développer une chaîne complète allant du texte brut à l'analyse syntaxique, et l'expérience a montré que les systèmes conservent ainsi des performances satisfaisantes. On renverra donc le lecteur aux actes de la tâche commune CoNLL 2018 sur ce point (Zeman *et al.*, 2018).

Conditions d'entraînement du système. Comme nous voulons explorer des scénarios pour des langues faiblement dotées, nous avons supposé que l'on ne disposait pas de données de développement permettant d'ajuster les paramètres du système (même dans le cas du same, pour lequel il existe maintenant des données importantes, notamment le corpus annoté syntaxiquement au format UD). Nous avons donc limité les expériences, notamment lors de la phase d'apprentissage, en considérant toutes les données disponibles une fois, sans arrêt anticipé, suivant Guo *et al.* (2016). D'autres stratégies seraient possibles (plusieurs itérations en faisant varier les phrases utilisées lors de l'apprentissage par exemple), mais le gain observé sur les résultats est minime et souvent non significatif. Ce type d'approches pose en outre des problèmes de réplicabilité et nous l'avons donc laissé de côté. Enfin, il faut noter que, pour l'élaboration d'un modèle multilingue, les différentes sources de données sont de taille très déséquilibrée. Pour pallier ce problème, et suivant les travaux antérieurs de Guo *et al.* (2016), nous avons effectué vingt fois plus d'itérations pour les langues faiblement dotées que pour les autres langues.

Comparaison avec la tâche commune CoNLL 2017. Nous avons utilisé les mêmes conditions pour l'entraînement de notre système dans les expériences décrites ici que pour la tâche commune CoNLL. En particulier nous n'avons pas de corpus de développement (nous disposons juste de vingt phrases annotées pour le same et de dix phrases annotées pour le komi pour la mise au point du système, comme dit plus haut).

Le tableau 2 présente les résultats obtenus par les trois meilleures équipes sur le same lors de la tâche commune CoNLL 2017, ainsi que les résultats de notre propre système. Il est évident, au vu de ces résultats, que notre système était alors loin d'être aussi performant que les meilleurs systèmes sur le same, à savoir ceux de Cornell (Shi *et al.*, 2017) , de Stuttgart (Björkelund *et al.*, 2017) ou de Harbin (Che *et al.*, 2017).

Lors de CoNLL 2017, C2L2 (Cornell Univ.) a obtenu les meilleures performances pour le same avec une approche par transfert délexicalisé (en utilisant un corpus de finnois pour l'entraînement et un corpus de développement de vingt phrases en same pour ajuster les paramètres du système, sans utilisation de traits lexicaux, c'est-à-dire en se fondant uniquement sur les étiquettes morphosyntaxiques et non sur les mots eux-mêmes). IMS (Stuttgart) a utilisé une approche similaire (approche par transfert délexicalisé) en utilisant pour l'entraînement quarante corpus différents encodés au format UD, et a ainsi obtenu le deuxième meilleur résultat.

Comparaison avec la tâche commune CoNLL 2018. Le same était à nouveau une langue de test lors de la campagne d'évaluation CoNLL 2018. Le meilleur système a obtenu les résultats suivants lors de la campagne 2018 : 69,87 LAS et 76,66 UAS. Ces résultats sont bien meilleurs que ceux rapportés pour l'évaluation CoNLL 2017 (tableau 2) ou même dans cet article (tableau 3), mais en 2018 des données d'entraînement importantes étaient fournies pour le same (il s'agissait essentiellement du corpus de same au format UD publié après la campagne d'évaluation 2017, comme indiqué dans l'introduction). Il est donc important de souligner que les résultats obtenus sur le corpus CoNLL 2018 ne sont en rien comparables avec les résultats 2017, où seules vingt phrases étaient disponibles pour la mise au point des systèmes.

Les résultats du meilleur système lors de la campagne d'évaluation CoNLL 2018 (69,87 LAS et 76,66 UAS) donnent malgré tout une idée de l'état de l'art pour une langue à morphologie riche, comme le same, quand on dispose d'un corpus d'entraînement moyennement volumineux. Ils permettent aussi d'avoir une idée de l'écart de performance entre une langue pour laquelle on dispose de données annotées et une langue pour laquelle on ne dispose pas de telles ressources (entre 10 et 15 points de différence environ), et aussi une idée de l'écart par rapport à l'anglais (là aussi, entre 10 et 15 points de différence – ces chiffres sont évidemment à prendre avec précaution car il faudrait faire d'autres expériences, avec d'autres langues et des conditions expérimentales plus directement comparables pour obtenir des résultats vraiment fiables). On confirme là, par l'observation, des résultats très évidents : une langue synthétique à morphologie riche est plus complexe à analyser qu'une langue analytique avec une complexité morphologique moindre, et un corpus d'entraînement de grande taille est aussi un facteur majeur d'amélioration des performances. Pour le reste, redisons-le : les résultats CoNLL 2018 ne sont en rien comparables aux résultats 2017 du fait des conditions expérimentales radicalement différentes pour le same lors de ces deux campagnes.

6. Résultats et analyse

Les résultats pour le same du nord sont donnés dans le tableau 3, et les résultats pour le komi-zyriène dans le tableau 4. Les résultats du tableau 3 diffèrent de ceux du tableau 2 car les expériences faites après la campagne CoNLL 2017 ont permis de mieux utiliser les vingt phrases fournies pour la mise au point du système et surtout de tester différentes combinaisons de langues afin d'identifier le modèle le plus per-

formant pour la tâche. On voit que le système de base est ainsi plus performant que le système officiel ayant participé à CoNLL 2017, même sans ressources extérieures ni modèle multilingue.

	Corpus utilisés	**Score LAS**	**Score UAS**
1	sme (20)	32,96	46,85
2	eng (12 217)	32,72	50,44
3	fin (12 543)	40,74	54,24
4	sme (20) + eng (12 217)	46,54	61,61
5	**sme (20) + fin (12 543)**	**51,54**	**63,06**

Tableau 3. *Évaluation de l'analyse du same du nord (sme) : scores LAS (*labeled attachment scores*) et UAS (*unlabeled attachment scores*), c'est-à-dire scores calculés en prenant en compte l'étiquette de la relation (score LAS, colonne de gauche), et sans elle (score UAS, colonne de droite). La première ligne sme (20) réfère à l'expérience utilisant uniquement sur les vingt phrases annotées de same disponibles pour l'entraînement. Les autres lignes montrent les résultats avec différentes combinaisons de corpus annotés : anglais (eng) et finnois (fin). Pour chaque corpus, le nombre de phrases utilisées est indiqué entre parenthèses.*

Globalement, les expériences que nous avons menées en utilisant le finnois comme source de connaissances (en particulier pour élaborer des plongements de mots bilingues) ont permis d'obtenir de meilleurs résultats qu'avec d'autres langues (on obtient de meilleurs résultats avec le finnois qu'avec l'anglais pour l'analyse syntaxique du same du nord, cf. tableau 3, et on obtient également de meilleurs résultats avec le finnois qu'avec l'anglais pour l'analyse du komi, cf. tableau 4 – le russe est toutefois plus performant pour analyser le komi que le finnois). Ceci semble indiquer d'une

	Corpus utilisés	**Score LAS**	**Score UAS**
1	kpv (10)	22,33	51,78
2	eng (12 217)	44,47	59,29
3	rus (3 850)	53,85	71,29
4	fin (12 543)	48,22	66,98
5	kpv (10) + eng (12 217)	50,47	66,23
6	kpv (10) + rus (3 850)	53,10	69,98
7	kpv (10) + fin (3 850)	53,66	71,29
8	kpv (10) + fin (12 543)	55,16	**73,73**
9	kpv (10) + eng (12 217) + fin (12,543)	52,50	68,57
10	**kpv (10) + rus (3 850) + fin (12 543)**	**56,66**	71,86

Tableau 4. *Évaluation de l'analyse syntaxique du komi. La première ligne kpv (10) réfère à l'expérience utilisant uniquement les dix phrases annotées de komi disponibles pour l'entraînement. Les autres lignes montrent les résultats avec différentes combinaisons de corpus annotés : anglais (eng), russe (rus), et finnois (fin). Pour chaque corpus, le nombre de phrases utilisées est indiqué entre parenthèses.*

part qu'il est possible d'inférer des connaissances linguistiques utiles pour l'analyse à partir d'une langue tierce et, d'autre part, que la typologie et la situation linguistique jouent un rôle (on obtient de meilleurs résultats sur le same ou le komi à partir du finnois ou éventuellement du russe qu'à partir de l'anglais, même si les données utilisées pour l'anglais sont largement plus volumineuses).

La stratégie de transfert de connaissances d'une langue vers l'autre a bien fonctionné pour le finnois vis-à-vis du same, ce qui peut sembler logique car le finnois et le same sont supposés relativement proches génétiquement parlant (mais n'importe quel locuteur pourra aussi dire à quel point ces langues sont éloignées : il n'y a pas d'intercompréhension, même limitée, et même le vocabulaire de base est très différent). Le transfert fonctionne bien aussi pour le komi, alors que le komi est supposé plus éloigné du finnois d'un point de vue génétique.

Concernant le komi, une des hypothèses que nous avions faites *a priori* était qu'un modèle élaboré à partir du russe pourrait aboutir à de meilleurs résultats qu'un modèle acquis à partir du finnois, car le russe a beaucoup « contaminé » le komi depuis plusieurs décennies (du fait de la situation linguistique et du bilinguisme de tous les locuteurs komis). Les résultats pour le russe sont, de fait, étonnants (ligne 3 du tableau 4). On obtient ainsi d'excellents résultats, qui surpassent même les résultats obtenus en ajoutant les dix phrases annotées de komi lors de l'apprentissage (ligne 6). Ceci est en fait sans doute dû à la proximité du russe et du komi, à la présence de cognats et surtout, aux conditions dans lesquelles sont faites ces expériences. L'apprentissage d'analyseurs avec si peu de données est donc possible, mais il faut garder à l'esprit que les résultats sont alors relativement instables (ceci pose d'ailleurs la question de la validité des résultats obtenus à partir d'échantillons aussi petits). Il est probable que, dans d'autres conditions expérimentales, les résultats obtenus avec le corpus russe seul seraient moins bons.

Si l'on fait abstraction des performances obtenues pour le komi à partir du corpus russe seul, nos résultats montrent que l'ajout de phrases annotées de la langue à analyser, même en petite quantité, peut améliorer significativement les résultats obtenus (lignes 5 à 10 ; on l'avait déjà vu lors de notre participation officielle à la tâche commune CoNLL 2017, où l'usage des vingt phrases fournies par défaut, et le fait d'en réserver une partie ou non comme corpus de développement, pouvaient avoir un effet très positif sur les résultats finaux). La taille des corpus bruts utilisés pour l'obtention des plongements de mots, et les ressources annexes utilisées pour l'acquisition de connaissances linguistiques jouent aussi un rôle important (on comparera ainsi les lignes 7 et 8, où le modèle est à chaque fois conçu à partir du finnois, mais avec deux corpus de tailles différentes). C'est logiquement le corpus le plus grand qui permet d'obtenir le modèle le plus performant. On peut aussi comparer les performances relatives obtenues avec des modèles élaborés à partir du russe, du finnois et de l'anglais : l'anglais, même avec un ensemble de phrases annotées très supérieur, n'est pas vraiment compétitif sur le komi. Comme on l'a dit, ces résultats sont toutefois fragiles et sont validés sur un corpus très petit. Il faut donc souhaiter davantage d'études sur des

langues variées, afin d'obtenir une meilleure vue des types de résultats possibles en fonction des langues, des ressources et des algorithmes utilisés.

Finalement, c'est le modèle élaboré à partir du finnois et du russe qui permet d'obtenir les meilleurs résultats (et non celui élaboré à partir de l'anglais, même si on dispose de plus de données pour l'anglais). Il semble bien que les langues choisies pour l'apprentissage jouent un rôle, et il est important de choisir celles-ci en fonction de critères linguistiques et typologiques.

On observe enfin que les scores UAS (c'est-à-dire sans tenir compte des étiquettes de relations syntaxiques) varient légèrement plus que les scores LAS (scores avec les étiquettes des relations syntaxiques), autrement dit les relations de base ont été trouvées et la racine correctement identifiée dans un certain nombre de cas, même quand les étiquettes des relations n'ont pas été attribuées correctement. Il est intéressant de noter que le modèle qui obtient le meilleur score UAS est le couple komi-finnois, même si d'autres combinaisons (avec l'ajout du russe notamment) permettent d'obtenir les meilleurs scores LAS.

7. Conclusion

Dans cet article, nous avons présenté une approche fondée sur l'utilisation de modèles multilingues, afin de fournir une analyse syntaxique pour des langues disposant de peu de ressources, et en particulier ne disposant pas de données annotées (à part quelques phrases utilisées comme amorçage, dix à vingt phrases dans les expériences menées ici et lors de la campagne CoNLL 2017). Nous avons montré que l'approche suivie était efficace dans le cas du same et du komi, même si les performances restent évidemment bien en deçà de ce que l'on peut obtenir avec un corpus d'entraînement de grande taille, comme en témoignent nos résultats pour le same lors de la campagne d'évaluation CoNLL 2018 (avec un corpus d'entraînement au format UD). Il s'agit toutefois, à notre avis, d'un cadre intéressant pour aider à produire des corpus annotés pour des langues peu dotées. Le cas du komi est à cet égard un cas d'étude intéressant, dans la mesure où il s'agit d'une langue avec peu de ressources, mais avec des locuteurs intéressés et demandeurs d'outils d'analyse automatique. Ce cadre pose toutefois un problème d'évaluation, en l'absence de données de référence (*gold standard*).

Nous avons observé que les modèles multilingues permettent généralement d'améliorer les performances par rapport à des modèles monolingues. Les langues génétiquement liées semblent être la meilleure source de connaissances (le finnois est ainsi efficace pour l'analyse du same comme du komi), mais la prise en compte de langues de contact semble aussi pertinente (ainsi le russe pour l'analyse du komi), de même que des langues pour lesquelles on dispose tout simplement de gros corpus (comme l'anglais). Une meilleure compréhension de l'apport réel de chaque langue au processus global serait intéressante pour permettre de définir une stratégie plus générale, et surtout reproductible, concernant le développement et l'utilisation de modèles multilingues pour l'analyse syntaxique.

Remerciements

Les auteurs remercient les trois relecteurs anonymes pour leurs suggestions, qui ont permis de largement améliorer cet article. Les travaux décrits ont été en partie effectués dans le cadre du projet LAKME, financé par l'université Paris Sciences et Lettres (IDEX PSL référence ANR-10-IDEX-0001-02). Cette recherche a aussi bénéficié du soutien d'un projet RGNF-CNRS entre le Lattice et l'université d'État des sciences humaines de Russie.

Annexe : contribution à l'élaboration d'un corpus arboré pour le komi

Afin de permettre l'évaluation de l'analyseur décrit dans cet article, un ensemble de phrases en komi ont été annotées au format UD. Ce corpus comprend actuellement environ trois cents phrases, et devrait en contenir mille prochainement. Une première version de ce corpus a été incluse dans la distribution officielle Universal Dependencies d'avril 2018[9]. Au-delà de la réalisation d'un nouveau corpus arboré, plusieurs points peuvent être soulignés, qui nous semblent relativement typiques du cas des langues de terrain et des langues minoritaires.

Les données disponibles sont de deux types très différents. On a d'un côté des sources écrites, parfois anciennes, écrites dans une langue relativement élaborée et littéraire, parfois très éloignée de la langue quotidienne. De l'autre, on dispose de corpus beaucoup plus modestes (en taille) correspondant à des enquêtes de terrain faites par des linguistes. Ce matériau est précieux, car directement issu de travaux linguistiques, il rend compte de la langue réellement utilisée par les locuteurs au quotidien, mais il pose plusieurs difficultés. Des difficultés matérielles d'abord, dans la mesure où ces données sont souvent encodées dans des formats particuliers, qui ne conviennent pas directement à un traitement automatique ; des difficultés liées à la taille des corpus existants, qui rendent difficile l'utilisation de techniques d'apprentissage artificiel par exemple. Enfin, ces corpus sont représentatifs de l'oral : ils posent donc des problèmes particuliers et les outils mis au point pour l'écrit ne sont pas très performants sur ce type de données.

Notre travail se situe dans le cadre d'un effort en cours en vue de fournir des données annotées pour un certain nombre de langues finno-ougriennes. Des corpus arborés sont déjà disponibles pour le finnois, l'estonien et le hongrois. L'année 2017 a vu l'émergence d'un corpus arboré important pour le same du nord (produit en grande partie automatiquement à partir des outils d'analyse mis au point à l'université de Tromsø et non entièrement vérifié manuellement). Un corpus arboré est actuellement en préparation pour l'erzya (langue mordve), ce qui permettra de couvrir à terme une partie non négligeable des langues finno-ougriennes, même si des efforts seront encore nécessaires pour les autres langues. Une tendance similaire est observée pour ce qui

9. Voir le site officiel de l'initiative Universal Dependencies : `http://universaldependencies.org`.

concerne la réalisation de corpus arborés à partir du résultat d'enquêtes de terrain. À notre connaissance, des projets existent par exemple pour le dargwa (langue du Caucase), le pnar (langue austro-asiatique) et le shipibo-konibo (langue du Pérou).

En pratique...

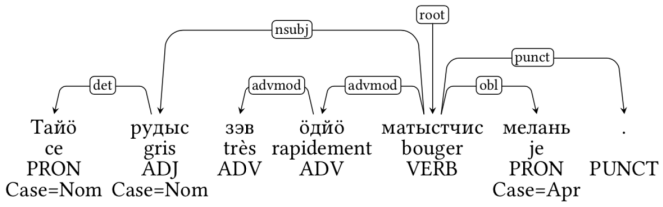
La plupart des travaux sur la langue komie sont actuellement menés à Syktyvkar, Russie (capitale de la République komie). Le FU-Lab, dirigé par Marina Fedina, a en particulier numérisé un nombre important de livres komis (datant du début du XX^{e} siècle jusqu'à aujourd'hui), ceux libres de droits étant directement mis à disposition en ligne. Ce corpus brut compte actuellement quarante millions de mots et l'objectif à long terme est de numériser tous les livres publiés en komi-zyriène. Le nombre total des publications est estimé à environ quatre mille cing cents livres, auxquels il faut ajouter des dizaines de milliers de pages issues de journaux et de revues. Pour la constitution de notre corpus, nous avons évidemment veillé à n'utiliser que des textes libres de droits, afin d'en assurer une distribution aussi large et simple que possible. Il est possible qu'à un stade ultérieur des matériaux moins standard puissent être inclus dans la base de données, par exemple des textes issus de blogs et de discussions en ligne, mais cela pose immédiatement des problèmes de droits et de diffusion.

En dehors du projet mené à Syktyvkar, un des plus grands projets de recherche sur le komi parlé a été un projet de documentation dirigé par Rogier Blokland (université de Uppsala) en 2014-2016. Ce projet a abouti à un grand corpus en langue parlée transcrite. Ces données sont précieuses, pour les raisons que nous avons données *supra*, mais elles sont aussi problématiques car elles ne peuvent généralement pas être diffusées directement dans le domaine public. Le corpus contient aussi des formes dialectales, et comme le komi-zyriène écrit ne suit pas les principes utilisés pour les transcriptions, il semble problématique de mélanger ce matériau avec les données issues de sources écrites. Le corpus oral contient enfin de nombreuses phrases où le komi est mêlé à du vocabulaire russe, les locuteurs pratiquant le code switching en permanence. Cette langue est donc non standard, mais elle est par ailleurs scientifiquement intéressante et pertinente.

À partir de ce point de départ, il a été décidé de créer deux corpus différents, le premier avec les matériaux écrits et le deuxième avec les données orales issues d'enquêtes de terrain.

Annotation syntaxique du corpus komi-zyriène au format UD

Pour l'annotation du corpus komi-zyriène, nous nous sommes inspirés de corpus arborés existants et des consignes d'annotation liées, notamment celles ayant été constituées pour le finnois, le same du nord et l'erzya, ainsi que pour le russe. Il s'agit de langues proches du komi (langues de la même famille, à l'exception du russe), et il nous semblait naturel d'aller voir du côté de ces langues en priorité. De fait, les



ID phrase: belyx-011.042
Traduction: Ce (nuage) gris est venu très rapidement vers moi.

consignes d'annotation ont facilement pu être transposées au komi et quasiment toutes les configurations observées correspondaient à des cas de figure observés (*mutatis mutandis*) dans au moins une de ces langues.

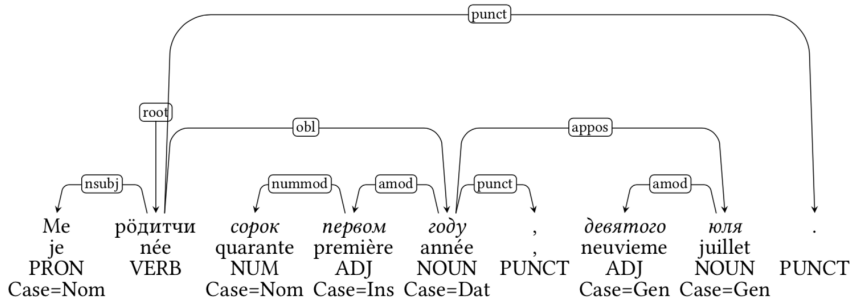
Le komi-zyriène présente malgré tout quelques particularités et différences qui le distinguent de ces langues proches. Il existe notamment deux cas spatiaux largement spécifiques au komi (en fait aux langues permiennes, branche des langues finno-ougriennes qui regroupe le komi et l'oudmourte) : l'égressif et l'approximatif. Ces deux cas expriment le mouvement depuis et vers une direction. Ils se distinguent de l'élatif et de l'illatif (deux cas bien répandus dans l'ensemble des langues finno-ougriennes) : l'élatif et l'illatif expriment aussi un mouvement depuis ou vers un lieu, mais ils insistent justement sur ce point de départ ou d'arrivée, alors que l'égressif et l'approximatif insistent davantage sur le mouvement, sans préciser le point de départ ou d'arrivée.

L'exemple ci-dessous illustre précisément l'utilisation du cas approximatif, sans aucun caractère égressif.

Il existe deux autres cas directionnels en komi, traditionnellement appelés prolatifs et transitifs, qui expriment le mouvement le long d'un chemin. Ceux-ci correspondent assez bien au cas appelé « perlatif » du guide d'annotation au format UD (Universal Dependencies), mais ce cas ne figure pas dans les exemples annotés jusqu'ici. Plus généralement, ceci pose la question de la terminologie employée et de la mise en rapport des cas (et plus généralement des notions linguistiques) à travers les langues : nous pensons que le prolatif et le translatif correspondent au perlatif, mais ceci mériterait sûrement une discussion approfondie. UD est en tout cas l'occasion de s'interroger sur la terminologie en cours, les notions manipulées et les correspondances entre langues. Il était à cet égard utile de garder un œil sur le russe lors de l'annotation, dans la mesure où le komi emprunte certaines constructions au russe. De plus, il semble souhaitable d'être, autant que faire se peut, cohérent et homogène au niveau des annotations.

Le premier exemple illustre une situation typique où la date est exprimée en russe, y compris au niveau du marquage morphologique et syntaxique.

Nous avons cherché ici à avoir une annotation comparable à celle de la structure correspondante en russe. Ce choix diffère de ce qui a été fait pour la plupart des autres



ID phrase: kpv_izva20150403-2-b.014
Traduction: Je suis née (l'année) 1941, le 9 juillet.

langues, où les mots ou structures en langue étrangère sont généralement marqués en tant que tels, et donc globalement plutôt mis de côté. Vu le bilinguisme de tous les locuteurs komis qui se retrouve en partie dans les corpus, nous souhaitons avoir une annotation qui intègre pleinement les passages en russe (y compris à l'intérieur d'une même phrase en cas de code switching comme ici) et les considère comme faisant pleinement partie du corpus komi.

Il faut toutefois noter que ceci peut aussi entraîner différents problèmes. Par exemple, certaines structures (provenant du russe) auront un trait `gender` (exprimant le genre grammatical), alors qu'il s'agit d'un trait morphologique étranger au komi. Ceci est évidemment aussi un défi pour les outils de TAL et les analyseurs en général, qui doivent gérer des situations linguistiques plus complexes que ce que l'on trouve dans la plupart des grands corpus monolingues disponibles. C'est cette richesse qui fait l'intérêt de ces langues trop longtemps laissées de côté.

8. Bibliographie

Aikio A., « An essay on Saami ethnolinguistic prehistory », *in* R. Grünthal, P. Kallio (eds), *A Linguistic Map of Prehistoric Northern Europe*, Société Finno-Ougrienne, Helsinki, p. 63-117, 2012.

Ammar W., Mulcaire G., Ballesteros M., Dyer C., Smith N. A., « Many Languages, One Parser », *Transactions of the Assoc. for Comp. Linguistics (TACL)*, vol. 4, p. 431-444, 2016a.

Ammar W., Mulcaire G., Tsvetkov Y., Lample G., Dyer C., Smith N. A., « Massively multilingual word embeddings », *Prépublication arXiv :1602.01925*, 2016b.

Artetxe M., Labaka G., Agirre E., « Learning principled bilingual mappings of word embeddings while preserving monolingual invariance », *Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Austin, p. 2289-2294, 2016.

Artetxe M., Labaka G., Agirre E., « Learning bilingual word embeddings with (almost) no bilingual data », *Conf. of the Assoc. for Comp. Linguistics (ACL)*, Vancouver, p. 451-462Proceedin, 2017.

Ballesteros M., Goldberg Y., Dyer C., Smith N. A., « Training with Exploration Improves a Greedy Stack LSTM Parser », *Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Austin, p. 2005-2010, 2016.

Björkelund A., Falenska A., Yu X., Kuhn J., « IMS at the CoNLL 2017 UD Shared Task : CRFs and Perceptrons Meet Neural Networks », *CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, p. 40-51, 2017.

Blokland R., Fedina M., Gerstenberger C., Partanen N., Rießler M., Wilbur J., « Language documentation meets language technology », *in* T. Pirinen, F. Tyers, T. Trosterud (eds), *Workshop on Comp. Linguistics for Uralic Languages*, Tromsø, 2015.

Bojanowski P., Grave E., Joulin A., Mikolov T., « Enriching Word Vectors with Subword Information », *Transactions of the Assoc. for Comp. Linguistics (TACL)*, vol. 5, p. 135-146, 2017.

Che W., Guo J., Wang Y., Zheng B., Zhao H., Liu Y., Teng D., Liu T., « The HIT-SCIR System for End-to-End Parsing of Universal Dependencies », *CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, p. 52-62, 2017.

Chen D., Manning C. D., « A Fast and Accurate Dependency Parser using Neural Networks. », *Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, p. 740-750, 2014.

Cho K., « Natural language understanding with distributed representation », *Prépublication arXiv :1511.07916*, 2015.

Conneau A., Lample G., Ranzato M., Denoyer L., Jégou H., « Word Translation Without Parallel Data », *Conf. International Conference on Learning Representations (ICLR)*, Toulon, 2017.

Das D., Petrov S., « Unsupervised Part-of-Speech Tagging with Bilingual Graph-Based Projections », *Conf. Assoc. for Comp. Linguistics (ACL)*, Portland, 2011.

Dozat T., Qi P., Manning C. D., « Stanford's Graph-based Neural Dependency Parser at the CoNLL 2017 Shared Task », *CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, p. 20-30, 2017.

Gerstenberger C., Partanen N., Rießler M., Wilbur J., « Utilizing language technology in the documentation of endangered Uralic languages », *Northern European Journal of Language Technology*, vol. 4, p. 29-47, 2016.

Gouws S., Bengio Y., Corrado G., « BilBOWA : Fast Bilingual Distributed Representations without Word Alignments », *Intern. Conf. on Machine Learning (ICML)*, Lille, p. 748-756, 2015.

Gouws S., Søgaard A., « Simple task-specific bilingual word embeddings », *Conf. of the North American Chapter of the Assoc. for Comp. Linguistics – Human Language Technologies (NAACL-HLT)*, Denver, p. 1386-1390, 2015.

Guo J., Che W., Yarowsky D., Wang H., Liu T., « Cross-lingual Dependency Parsing Based on Distributed Representations », *Conf. of the Assoc. for Comp. Linguistics (ACL)*, Beijing, 2015.

Guo J., Che W., Yarowsky D., Wang H., Liu T., « A Representation Learning Framework for Multi-Source Transfer Parsing. », *Conf. of the Association for the Advancement of Artificial Intelligence (AAAI)*, Beijing, p. 2734-2740, 2016.

Huang Z., Xu W., Yu K., « Bidirectional LSTM-CRF Models for Sequence Tagging », *Prépublication arxiv1508.01991*, 2015.

Hwa R., Resnik P., Weinberg A., Cabezas C., Kolak O., « Bootstrapping Parsers via Syntactic Projection Across Parallel Texts », *Natural Language Engineering*, vol. 11, nº 3, p. 311-325, 2005.

Khapra M. M., Sohoney S., Kulkarni A., Bhattacharyya P., « Value for Money : Balancing Annotation Effort, Lexicon Building and Accuracy for Multilingual WSD », *Coling*, Beijing, p. 555-563, 2010.

Kim S., Jeong M., Lee J., Lee G. G., « A Cross-lingual Annotation Projection Approach for Relation Detection », *Coling*, Beijing, p. 564-571, 2010.

Kim S., Jeong M., Lee J., Lee G. G., « Cross-Lingual Annotation Projection for Weakly-Supervised Relation Extraction », *ACM Transactions on Asian Language Information Proc. (TALIP)*, vol. 13, nº 1, p. 3 :1-3 :26, February, 2014.

Kiperwasser E., Goldberg Y., « Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representations », *Transactions of the Assoc. for Comp. Linguistics (TACL)*, vol. 4, p. 313-327, 2016.

Kozhevnikov M., Titov I., « Cross-lingual bootstrapping for semantic role labeling », *xLiTe : Cross-Lingual Technologies*, Lake Tahoe, 2012.

Leinonen M., « The Russification of Komi », *The Slavicization of the Russian North. Mechanisms and Chronology*, Slavica Helsingiensia 27, p. 234-245, 2006.

Levy O., Søgaard A., Goldberg Y., « A Strong Baseline for Learning Cross-Lingual Word Embeddings from Sentence Alignments », *Conf. of the European Chapter of the Assoc. for Comp. Linguistics (EACL)*, Valence, p. 765-774, 2017.

Lim K., Partanen N., Poibeau T., « Multilingual Dependency Parsing for Low-Resource Languages : Case Studies on North Saami and Komi-Zyrian », *Language Resource and Evaluation Conference (LREC)*, Miyazaki, 2018.

Lim K., Poibeau T., « A System for Multilingual Dependency Parsing based on Bidirectional LSTM Feature Representations », *CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, p. 63-70, 2017.

Liu K., Lü Y., Jiang W., Liu Q., « Bilingually-Guided Monolingual Dependency Grammar Induction », *Conf. of the Assoc. for Comp. Linguistics (ACL)*, Sofia, p. 1063-1072, 2013.

Luong T., Pham H., Manning C. D., « Bilingual Word Representations with Monolingual Quality in Mind. », *Conf. of the North American Chapter of the Assoc. for Comp. Linguistics – Human Language Technologies (NAACL-HLT)*, Denver, p. 151-159, 2015.

McDonald R., Crammer K., Pereira F., « Online large-margin training of dependency parsers », *Conf. of the Assoc. for Comp. Linguistics (ACL)*, Ann Arbor, p. 91-98, 2005a.

McDonald R., Pereira F., Ribarov K., Hajič J., « Non-projective Dependency Parsing Using Spanning Tree Algorithms », *Conf. on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP)*, Stroudsburg, p. 523-530, 2005b.

Mikolov T., Chen K., Corrado G., Dean J., « Efficient Estimation of Word Representations in Vector Space », *Prépublication arXiv :1301.3781*, 2013a.

Mikolov T., Le Q. V., Sutskever I., « Exploiting similarities among languages for machine translation », *Prépublication arXiv :1309.4168*, 2013b.

Naseem T., Barzilay R., Globerson A., « Selective sharing for multilingual dependency parsing », *Conf. of the Assoc. for Comp. Linguistics (ACL)*, Bruxelles, p. 629-637, 2012.

Nivre J., « Incrementality in deterministic dependency parsing », *Workshop on Incremental Parsing (organisé avec la conf. ACL 2004)*, Barcelone, p. 50-57, 2004.

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., « UDPipe : Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing », *Language Resources and Evaluation Conf. (LREC)*, Portorož, 2016.

Ruder S., Vulić I., Søgaard A., « A survey of cross-lingual embedding models », *Prépublication arXiv :1706.04902*, 2017.

Scherrer Y., Sagot B., « A language-independent and fully unsupervised approach to lexicon induction and part-of-speech tagging for closely related languages », *Language Resources and Evaluation Conf. (LREC)*, Reykjavik, 2014.

Shi T., Wu F. G., Chen X., Cheng Y., « Combining Global Models for Parsing Universal Dependencies », *CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, p. 31-39, 2017.

Smith D. A., Eisner J., « Parser Adaptation and Projection with Quasi-synchronous Grammar Features », *Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Singapour, p. 822-831, 2009.

Straka M., Hajič J., Straková J., « UDPipe : Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing », *Language Resources and Evaluation Conf. (LREC)*, Portorož, 2016.

Täckström O., McDonald R., Uszkoreit J., « Cross-lingual word clusters for direct transfer of linguistic structure », *Conf. of the North American chapter of the Assoc. for Comp. Linguistics (NAACL)*, Montréal, p. 477-487, 2012.

Taskar B., Chatalbashev V., Koller D., Guestrin C., « Learning structured prediction models : A large margin approach », *Int. Conf. on Machine Learning (ICML)*, Bonn, p. 896-903, 2005.

Vulić I., Søgaard A., Ruder S., « On the Limitations of Unsupervised Bilingual Dictionary Induction », *Conf. of the Assoc. for Comp. Linguistics (ACL)*, Melbourne, 2018.

Weiss D., Alberti C., Collins M., Petrov S., « Structured Training for Neural Network Transition-Based Parsing », *Conf. of the Assoc. for Comp. Linguistics (ACL)*, Beijing, 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 », *CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, Bruxelles, p. 1-20, 2018.

Zeman, D. *et al.*, « CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies », *CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, p. 1-19, 2017.

Zhao H., Song Y., Kit C., Zhou G., « Cross Language Dependency Parsing Using a Bilingual Lexicon », *Conf. of the Assoc. for Comp. Linguistics (ACL)*, p. 55-63, 2009.

Zhuang T., Zong C., « Joint Inference for Bilingual Semantic Role Labeling », *Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Cambridge, USA, p. 304-314, 2010.

Un état de l'art du traitement automatique du dialecte tunisien

Jihene Younes[1]* — **Emna Souissi**[2]** — **Hadhemi Achour**[3]* — **Ahmed Ferchichi**[4]*

* *Université de Tunis, ISGT, LR99ES04 BESTMOD, 2000, Le Bardo, Tunisia*
1 *jihene.younes@gmail.com,*
3 *Hadhemi_Achour@yahoo.fr*
4 *Ahmed.Ferchichi@planet.tn*
** *Université de Tunis, ENSIT, 1008, Montfleury, Tunisia*
2 *emna.souissi@ensit.rnu.tn*

RÉSUMÉ. *Dans le domaine du traitement automatique de la langue arabe, la majorité des recherches menées et des réalisations accomplies ont porté principalement sur l'arabe standard moderne (ASM). Les divers dialectes arabes (DA) comptent encore parmi les langues sous-dotées. Ce n'est que depuis une dizaine d'années que ces dialectes ont commencé à susciter un intérêt accru au sein de la communauté TAL, notamment compte tenu de leur utilisation de plus en plus importante sur le Web social. Dans ce travail, nous nous focalisons sur le dialecte tunisien (DT), et proposons de fournir un état de l'art sur le traitement automatique de ce dialecte. Une revue des travaux accomplis à ce jour, ainsi qu'un inventaire détaillé des divers outils TAL et ressources linguistiques disponibles pour le DT sont présentés puis discutés.*

ABSTRACT. *In the area of Arabic Natural Language Processing, most of the undertaken research and achievements have mainly involved Modern Standard Arabic (MSA). The various Arabic dialects (AD) are still considered to be among under-resourced languages. It's only in the last decade that these dialects began to arouse the interest of NLP researchers, especially given their increasing use on the social web. In this work, we focus on the Tunisian dialect (TD), and propose to provide a state of the art on the automatic processing of this dialect. A review of the works carried out to date and a detailed inventory of the various NLP tools and language resources available for the TD are presented and discussed.*

MOTS-CLÉS : *dialecte tunisien*$_1$*, ressources linguistiques*$_2$*, traitement automatique des langues*$_3$.

KEYWORDS: *Tunisian dialect*$_1$*, Language resources*$_2$*, Natural language processing*$_3$.

1. Introduction

La diglossie est l'une des principales caractéristiques de la langue arabe. Dans les pays arabes, il existe au moins deux formes d'arabe qui coexistent : l'arabe standard moderne (ASM) d'une part, utilisé comme langue officielle, et l'arabe dialectal d'autre part, recouvrant divers dialectes régionaux. Les dialectes arabes (DA) sont des variantes régionales de la langue arabe, naturellement parlées par les populations arabes et utilisées dans leur communication quotidienne. Ils diffèrent d'une région à une autre et d'un pays à un autre et parfois même, différents dialectes peuvent exister dans un même pays.

Les DA sont souvent considérés par la communauté TAL comme faisant partie des langues sous-dotées (Novotney *et al.*, 2016 ; Harrat *et al.*, 2017) compte tenu de la rareté des ressources linguistiques et outils TAL disponibles associés à ces dialectes. En effet, dans le domaine du traitement automatique de la langue arabe, la majorité des recherches menées et des réalisations accomplies ont porté principalement sur l'ASM. Étant la langue utilisée dans la presse, documents administratifs et officiels et enseignée dans les écoles, les ressources linguistiques électroniques en ASM sont largement répandues et disponibles, contrairement aux DA qui sont des langues essentiellement parlées et très rarement écrites. La rareté de ressources volumineuses écrites en arabe dialectal a constitué probablement l'un des freins majeurs à son étude et à son traitement automatique.

Par ailleurs, les DA commencent depuis une dizaine d'années à susciter un intérêt croissant au sein de la communauté TAL, s'expliquant, entre autres, par la prolifération de divers contenus textuels informels sur le Web et plus particulièrement sur le Web social, et qui sont de plus en plus riches en arabe dialectal. Le Web devient, ainsi, une source privilégiée par plusieurs chercheurs, pour l'extraction de contenus dialectaux et la construction de diverses ressources linguistiques. En parallèle, plusieurs travaux portant sur les DA visant, entre autres, leur analyse morphosyntaxique, le traitement de la parole ou encore l'analyse des sentiments ont vu le jour ces dernières années.

Dans ce travail, nous nous penchons sur le dialecte tunisien en particulier, et nous proposons d'élaborer un état de l'art sur son traitement automatique. Dans cette étude, une revue des travaux accomplis à ce jour avec des tableaux synthétiques indiquant les tailles des ressources, les approches utilisées et les performances, ainsi qu'un inventaire détaillé des divers outils TAL et des ressources linguistiques disponibles pour le DT sont présentés puis discutés. Notons, à ce niveau, que notre investigation a porté sur tous les travaux publiés avant la fin du mois de mars 2018. Nous ne rapportons, pour les travaux qui ont porté sur un groupe de dialectes arabes comprenant le tunisien[1], que ceux qui mentionnent les tailles des données utilisées et les performances

1. Parmi les travaux ayant porté sur un groupe de dialectes comprenant le tunisien, sans préciser les tailles de données ni les performances obtenues spécifiques à ce dialecte, nous pouvons citer (Suwaileh *et al.*, 2016 ; Eldesouki *et al.*, 2017).

obtenues explicitement pour le DT. Ce travail pourrait ainsi constituer un point de départ pour tout chercheur souhaitant travailler sur le traitement automatique du DT.

2. Dialecte tunisien (DT)

La Tunisie compte plus de 11 millions d'habitants[2] dans la région du Maghreb située en Afrique du Nord, avec une diaspora de plus de 1,3 million de personnes[3]. La langue maternelle des Tunisiens est le dialecte tunisien, souvent appelé par ses locuteurs « darija » (en arabe, « دارجة », qui est le féminin de la forme « دارج » qui signifie « familier, commun, populaire, utilisé »[4]) . Il est aussi courant de se référer au DT directement en tant que langue nommée « tounsi » (tunisien) .

Le DT est principalement une langue parlée, spontanément utilisée par les Tunisiens pour leur communication de tous les jours, mais il est parfois aussi un langage littéraire au moyen duquel on dit des proverbes, des comptines, des contes, des devinettes et des poèmes et une langue d'écriture lorsque par exemple les paroles de chansons et de pièces de théâtre sont écrites en DT (Pereira, 2005). Aujourd'hui, il est largement utilisé à la radio, à la télévision et dans la publicité. Étant une variante de la langue arabe, le DT est en général, et depuis longtemps, écrit à l'aide de l'alphabet arabe, mais il peut aussi être écrit en utilisant l'alphabet latin. Durant ces dernières années, l'adoption massive de nouveaux modes de communication (SMS, e-mails, Facebook, Twitter, etc.) a grandement renforcé l'écriture de messages et de divers contenus textuels en dialecte, notamment dans le script latin.

2.1. *Spécificités linguistiques*

Le dialecte tunisien est un dialecte arabe maghrébin qui présente d'importantes dissimilarités avec l'ASM d'une part, et avec les autres dialectes arabes d'autre part, surtout avec les dialectes orientaux (arabe Mashriqi). Le DT est en effet, très peu compris par les arabophones orientaux (d'Égypte, du Soudan, du Levant, d'Irak et de la péninsule Arabique) car il dérive de différents substrats et d'un mélange de plusieurs langues (Sayahi, 2014 ; Mohand, 1999 ; Elimam, 2009 ; Elimam, 2012 ; Mzoughi, 2015) : punique, berbère, arabe, turc, français, espagnol et italien. Les différentes civilisations qui ont transité par le pays, ont en effet, laissé leurs empreintes dans le patrimoine linguistique tunisien et le dialecte parlé par les Tunisiens est aujourd'hui bien riche en emprunts lexicaux d'origines multiples. Nous pouvons citer, à titre d'exemples les mots du DT: « كرموس - karmous », qui signifie « figue » et a pour origine le mot berbère « takarmoust », « فرفطو - farfatttou », qui signifie « papillon » et

2. Selon http://countrymeters.info, consulté le 10 janvier 2018.
3. Selon http://ote.nat.tn/wp-content/uploads/2018/05/Repartition_de_la_communaute_tunisienne_a_l_etranger__2012.pdf, consulté en juillet 2018.
4. Dans le dictionnaire « معجم المعاني عربي عربي » : www.almaany.com

a pour origine le mot italien « farfalla », « برجوازي - borjouazi », qui signifie « bourgeois » et a pour origine le mot français « bourgeois », « صباط - sabbat », qui signifie « chaussure » et a pour origine le mot espagnol « zapato », etc. De plus, le DT est une langue en continuelle évolution dans la mesure où de nouveaux mots étrangers sont fréquemment intégrés dans le dialecte et souvent conjugués selon les règles de l'arabe (Sayahi, 2014 ; Elimam, 2009). Baccouche (1994) appelle ce phénomène, pour le cas du français, la « tendance de *tunisifier* le français ». Les exemples suivants montrent la dynamique du système linguistique du dialecte tunisien : « ريفز - rivez » (« réviser »), « يريفز - yrivez » (« il révise »), « مريفز - mrivez » (« en situation de réviser »), etc.

Il est aussi à noter que sur le plan lexical, le DT comprend des mots qui lui sont bien spécifiques et qui le distinguent des autres dialectes maghrébins et orientaux, tels que : « باهي - bahi » (« d'accord »), « برشة - barcha » (« beaucoup »), « فيسع - fissaa » (« vite »), « عالسلامة - asslama » (« bonjour »), « بالسلامة - bislama » (« au revoir »), « ينجّم - ynajjim » (« il peut »), « يزّي - yezzi » (« Arrête, ça suffit ! »), etc.

Sur le plan phonologique, certains phonèmes non arabes peuvent être utilisés dans le DT, tels que / g / ڤ, / p / پ et / v / ڤ. Les voyelles longues sont généralement raccourcies et les trois voyelles courtes sont souvent réduites à deux. Dans de nombreux cas, CvCC est changé en CCvC (par exemple, « سقف / saqf » (« toit ») en ASM, est dit « سقف / sqaf » en DT). En outre, la préférence est pour le stress de syllabe finale, en particulier avec la réduction des voyelles courtes non stressées (par exemple, « كتاب / kitaab » (« livre ») en ASM, est dit « كتاب / ktaab » en DT).

Sur le plan morphologique, plusieurs aspects distinguent le DT de l'arabe standard. En effet, dans le DT, la catégorie du nombre duel de l'ASM n'existe pas et les marques casuelles nominales sont aussi supprimées. La conjugaison des verbes en DT présente aussi des dissimilarités avec l'ASM, en utilisant des affixes différents. Le tableau 1 montre quelques exemples de caractéristiques spécifiques au DT concernant la conjugaison des verbes.

Conjugaison des verbes en DT	Exemple
Le préfixe n-ن est utilisé avec la première personne du singulier	نـ كتب / **n**iktib (j'écris)
Le préfixe n-ن avec le suffixe u-و sont utilisés pour le pluriel	نـ كتبـ و / **n**iktb**ou** (nous écrivons)
Le futur utilise le préfixe verbal (باش, بش) plus le verbe	**باش** نكتب / **bech** niktib (je vais écrire)

Tableau 1. *Conjugaison des verbes en DT*

Au niveau syntaxique, la structure de la phrase en DT présente des différences par rapport à l'ASM et parfois, les caractéristiques syntaxiques du système dialectal

tunisien sont en rupture complète avec l'ASM. Par exemple, la structure la plus dominante pour une phrase en DT est (SVO : Sujet Verbe Objet) tandis que pour l'ASM la structure la plus dominante est (VSO) (Saidi, 2014). En DT, un seul pronom relatif (« اللّي-elli », « qui ») remplace tous les autres pronoms relatifs de l'ASM. De même, les pronoms personnels sont réduits à 7 pronoms par opposition aux douze pour l'ASM, la forme duelle des pronoms démonstratifs est abandonnée, etc. (Mejri *et al.*, 2009).

2.2. *Présence sur les réseaux sociaux*

La Tunisie a connu ces dernières années une utilisation accrue des médias sociaux. Si l'on prend comme exemple Facebook, qui constitue aujourd'hui la plateforme de médias sociaux la plus populaire en Tunisie, le nombre total d'utilisateurs au cours des trois dernières années est passé de 4,6 en 2014 à 6,1 millions d'utilisateurs en 2017 (Mourtada *et al.*, 2014 ; Salem, 2017), ce qui représente plus de 53 % de la population tunisienne.

Pour ce qui est des langues utilisées sur le Web social, il est important de noter que les contenus générés par les utilisateurs tunisiens sont fortement multilingues, comprenant principalement les trois langues : arabe, français et anglais. Selon les mêmes études (celles de Mourtada *et al.* (2014), Salem (2017)), l'utilisation de la langue arabe sur le réseau Facebook en Tunisie a connu une croissance significative. Le taux d'utilisateurs du réseau utilisant la langue arabe est passé en effet de 18 % en 2016 à 68,6 % en 2017, de 15 % à 19,5 % pour l'anglais et de 91 % à 93,1 % pour le français. Remarquons ici, que malgré la forte croissance de l'utilisation de la langue arabe, le français reste une langue dominante sur le réseau social Facebook. La langue arabe utilisée sur le Web social ne se limite pas à l'ASM uniquement, mais les contenus textuels exprimés en arabe dialectal prolifèrent également sur les plateformes sociales. De plus, les productions dialectales peuvent être écrites en utilisant à la fois l'alphabet arabe et l'alphabet latin. Dans (Younes *et al.*, 2015), une étude menée sur différentes pages Facebook tunisiennes (politique, média, sport, etc.), a montré que 58 % du contenu total étudié est en dialectal et que 81,3 % des productions en DT sont transcrites en alphabet latin. Cela est expliqué dans (Younes *et al.*, 2015) par plusieurs facteurs comme la rareté des claviers arabes au début des années du Web et de la technologie mobile ainsi que le phénomène du multilinguisme caractérisant la population tunisienne.

2.3. *Difficultés et défis*

Le traitement automatique du DT est une tâche non triviale présentant des difficultés particulières. Ces difficultés sont principalement dues au manque de conventions de normes orthographiques, à l'évolution continue de la langue avec l'emprunt de nouveaux mots, aux variétés du DT pouvant différer d'une région à une autre à travers le pays et présenter des dissemblances à différents niveaux linguistiques (Baccouche et Mejri, 2004), ainsi qu'à ses importantes dissimilarités avec l'ASM. Ces défis rendent

difficile l'utilisation directe des outils de TAL disponibles pour l'ASM. Notons aussi que les phénomènes de diglossie et alternance de code (*code switching*) posent, notamment, le problème de l'identification du DT. Cette identification constitue une tâche complexe qui doit résoudre de nombreux problèmes d'ambiguïté. L'ambiguïté concerne les mots pouvant être à la fois des mots de l'ASM et du DT lorsqu'ils sont transcrits en alphabet arabe, ou encore les mots pouvant être à la fois des mots étrangers (français ou anglais) et du DT, lorsqu'ils sont transcrits en alphabet latin. Par exemple, le mot « خاطر » signifie « parce que » en DT et « esprit » en ASM, le mot « bard » signifie « froid » en DT et « poète » en anglais, le mot « flous » signifie « argent » en DT et « flous » en français, etc. (Younes *et al.*, 2015). Ces divers problèmes accentuent la difficulté du traitement automatique du DT, qui ne pourra pas se développer sans la disponibilité de larges ressources linguistiques et le développement de divers outils de TAL appropriés. Dans la suite de cet article, nous proposons une revue des travaux de TAL portant sur le DT et dressons un inventaire des outils et ressources actuellement disponibles pour son traitement automatique.

3. Traitement automatique du DT

3.1. *Construction de ressources linguistiques*

Dans cet article, nous nommons « ressource linguistique » (RL), toute collection de données relatives à la langue telles que les corpus oraux ou écrits, lexiques, dictionnaires, ontologies, etc. Ce type de ressources représente, en effet, un matériau essentiel pour l'étude des langues et le développement d'outils et applications de TAL. Pour ce qui concerne le DT, et vu le manque de ressources disponibles dans un format électronique pour ce dialecte, plusieurs travaux ont porté sur la construction de divers types de collections de données dialectales. Certains d'entre eux sont partis d'enregistrements de dialogues, conversations, radios et télédiffusions comme (Belgacem, 2009 ; Graja *et al.*, 2010 ; Masmoudi *et al.*, 2014a ; Masmoudi *et al.*, 2014b ; Masmoudi *et al.*, 2014c) et ont procédé à leur transcription en corpus écrits. D'autres chercheurs ont eu recours, entre autres, au Web et aux médias sociaux afin d'extraire des données en dialectal et construire divers types de ressources. Nous pouvons citer parmi ceux-ci, McNeil et Faiza (2011) qui ont construit un corpus appelé TAC (*Tunisian Arabic Corpus*) dans le cadre d'un projet de création d'un dictionnaire DT-anglais. Ce corpus a ensuite été organisé dans une application Web permettant un traitement linguistique de base. Younes *et al.* (2014 ; 2015) ont eu recours au Web social pour construire diverses ressources pour le DT. Ils ont développé un module qui extrait automatiquement les commentaires des pages Facebook tunisiennes et filtre les spams et les messages écrits en langues étrangères en s'appuyant sur des lexiques.

Mise à part la génération de corpus bruts pour le DT, il convient de mentionner les travaux qui ont visé la création de corpus parallèles et divers corpus annotés. Nous citons à cet effet, Bouamor *et al.* (2014) et Harrat *et al.* (2017). Par ailleurs, divers

corpus annotés ont été construits, notamment par Graja *et al.* (2013) et Zribi *et al.* (2015) et décrits dans les tableaux 2 et 3.

Les travaux de Boujelbane (2013) et Boujelbane *et al.* (2013a ; 2013b ; 2014) ont conduit à la construction de lexiques bilingues ASM-DT, en utilisant le Penn Arabic Treebank (Maamouri *et al.*, 2014) et des règles de conversion fondées sur les différences entre l'ASM et le DT. Hamdi *et al.* (2014) ont construit un lexique pour les noms déverbaux en DT. Masmoudi *et al.* (2014a ; 2014b ; 2014c), qui ont construit un corpus nommé TARIC (corpus arabe d'interaction du chemin de fer tunisien) en transcrivant manuellement des enregistrements audio, ont généré automatiquement un dictionnaire de prononciation du DT nommé TunDPDic.

Nous avons pu identifier, par ailleurs, des ontologies de domaine construites dans le cadre de travaux sur le traitement du DT dans les systèmes de dialogue (dans les gares) par Graja *et al.* (2011a ; 2011b), Karoui *et al.* (2013a ; 2013b) et Graja *et al.* (2015). En effet, Graja *et al.* (2011a ; 2011b) ont utilisé des ontologies pour couvrir le lexique utilisé dans les gares en utilisant le corpus TuDiCoI construit par Graja *et al.* (2010). Karoui *et al.* (2013a ; 2013b) ont proposé une méthode hybride (combinant une approche statistique et une approche linguistique) pour la construction semi-automatique d'une ontologie de domaine, appelée Railway Information Ontology (RIO), à partir du corpus TuDiCoI (Graja *et al.*, 2010). Des travaux sur la construction d'ontologies en DT comprennent également le WordNet (TunDiaWN) proposé par Bouchlaghem *et al.* (2014) et qui peut être considéré comme une ressource parallèle DT-ASM puisqu'il préserve le contenu AWN (Arabic WordNet de Elkateb *et al.* (2006)). À partir d'un corpus nommé MultiTD (Multi-source Tunisian Dialect corpus), Bouchlagem *et al.* (2014) ont utilisé une méthode permettant de regrouper les mots en groupes significatifs, fondée sur l'algorithme de K-modes (Huang, 1998) pour enrichir le WordNet. Citons enfin, l'ontologie « aebWordNet », proposée par Ben Moussa Karmani *et al.* (2014 ; 2015) et Ben Moussa Karmani et Alimi (2016), qui a été modélisée à partir du dictionnaire bilingue anglais-arabe tunisien « *Peace corps dictionary* » de Abdelkader (1977). Une synthèse de ces travaux, résumant les caractéristiques des ressources construites, est présentée dans les tableaux 2 et 3.

3.2. *Traitement de la parole*

Parmi les travaux de recherche menés sur le dialecte tunisien, nous pouvons distinguer ceux qui ont porté sur le traitement de la parole. Certains d'entre eux ont visé l'identification de différents dialectes arabes (comprenant le tunisien). Dans cette catégorie, nous citons Belgacem *et al.* (2010) qui ont utilisé le modèle de mélange gaussien (GMM) de Reynolds *et al.* (2000), qui permet de reconnaître les similitudes et les différences entre chaque dialecte. Les travaux de Lachachi et Adla (2015 ; 2016a ; 2016b) ont aussi porté sur deux systèmes d'identification automatique de 5 dialectes arabes du Maghreb : marocain, tunisien et trois dialectes algériens. Le premier système est basé sur les modèles de mélange de lois gaussiennes (GMM) et le second est fondé sur une combinaison d'un modèle du monde et des modèles de mélanges de lois gaussiennes

Auteurs	ADA	Script	Ressources construites
Belgacem (2009)	✓	A	- Corpus arabe multidialectal, de 10 h de parole de 8 DA dont 37 % transcrites (90 min dont 5 % transcrites pour le DT). - Source : discours enregistrés, journaux radio ou télédiffusés + transcription
Graja *et al.* (2010)		A	- Corpus TuDiCoI : 127 dialogues ; 893 segments ; 3 403 mots. - Source : conversations enregistrées dans la gare de la SNCFT + transcription
McNeil et Faiza (2011) ; McNeil (2015)		A	- Corpus TAC 2011 : 400 k mots ; Corpus TAC 2015 : 820 k mots - Source : écrits traditionnels + blogs + e-mails + Facebook + audio transcription
Graja *et al.* (2011a ; 2011b)		A	- Ontologie de 15 concepts - Source : corpus TuDiCoI
Karoui *et al.* (2013a ; 2013b)		A	- RIO : 14 concepts, 25 relations, 387 instances - Source : corpus TuDiCoI
Graja *et al.* (2013)		A	- Corpus TuDiCoI : 1 825 dialogues ; 12 182 segments ; 21 682 mots dont 7 814 mots sont annotés (normalisation lexicale, analyse morphologique, lemmatisation et attribution des synonymes) - Source : conversations enregistrées dans la gare de la SNCFT + transcription
Boujelbane *et al.* (2013a ; 2013b ; 2014b) ; Boujelbane (2013)		A	- Corpus : 5 h 20 min de paroles transcrites ; 37 964 mots dont 12 149 en DT - Corpus : 12 k mots - Lexique bilingue ASM-DT - Source: Arabic Tree Bank (ATB) + paroles transcrites + numérisation de la constitution tunisienne
Bouamor *et al.* (2014)	✓	A	- Corpus parallèle multidialectal MPCA : 2 000 phrases pour chaque langue dont 10 896 mots pour le DT (5 DA + ASM + anglais). - Source : 2 000 phrases du corpus égyptien-anglais de Zbib et Callison-Burch (2012)

Tableau 2. *Construction de ressources linguistiques en DT (1) (ADA : avec d'autres dialectes arabes : nous marquons dans cette colonne les travaux qui ne sont pas spécifiques au DT, mais qui l'ont traité comme faisant partie d'un ensemble de dialectes arabes (DA), script : le système d'écriture des ressources (A : arabe / L : latin))*

(UBM-GMM). Ils ont collecté un corpus de dialectes parlés à partir d'émissions télévisées. Pour le tunisien, le corpus comprend 53,37 heures avec 130 locuteurs.

D'autres travaux ayant porté sur la reconnaissance automatique de la parole en DT comprennent ceux de Neifar *et al.* (2014), fondés sur l'adaptation d'un système élaboré pour l'ASM, les travaux de Hassine *et al.* (2016) qui ont développé un système de reconnaissance automatique de la parole en arabe afin de reconnaître 10 chiffres arabes (de 0 à 9) parlés en dialectes marocain et tunisien ainsi que ceux de (Hassine *et al.*, 2018) portant sur la reconnaissance de la parole en DT. La conversion de Graphème-en-Phonème (G2P) pour le DT, consistant à convertir une séquence de graphèmes en une séquence de symboles phonétiques, a fait l'objet des travaux de Masmoudi *et al.* (2016) ainsi que Masmoudi *et al.* (2017). Tous ces travaux sont décrits dans le tableau 4.

Auteurs	ADA	Script	Ressources construites
Younes et Souissi (2014)		L	- Corpus de 43 222 messages en DT - Lexique DT de 19 763 mots - Source : SMS + Chat + forum + Facebook
Hamdi *et al.* (2014)		A	- Lexique bilingue ASM-DT de 39 793 entrées (14 804 - 5 017 lemmes) - Source : Arabic Tree Bank (ATB) (Maamouri *et al.*, 2014)
Masmoudi *et al.* (2014a ; 2014b ; 2014c)		A	- Corpus TARIC : 20 h de paroles transcrites ; 71 684 mots. - Dictionnaire phonétique TunDPDic : 18 k mots - Source: discours enregistré avec transcription manuelle
Bouchlaghem *et al.* (2014)		AL	- Corpus MultiTD de 32 848 mots - WordNet TunDiaWN - Types d'entités: *synset*, *word*, *form*, *word relations*, *annotator* - Source: corpus multiTD issu de réseaux sociaux (Twitter, Facebook, etc.), pièces de théâtre écrites, dictionnaires, discours enregistré, etc.
Ben Moussa Karmani *et al.* (2014 ; 2015) Ben Moussa Karmani et Alimi (2016)		AL	- WordNet (aebWordNet) - Synset : 18 209 entrées (8 279 lemmes et 25 748 mot-sens) - Source : dictionnaire bilingue anglais-DT de « *Peace corpus dictionary* »
Zribi *et al.* (2015)		A	- Corpus annoté STAC : 4 h 50 min - 42 388 mots – 7 788 phrases - Corpus écrit : annoté par segmentation en phrases, segmentation des mots en affixes et clitiques, lemme, genre, nombre, personne, voix, étiquettes, etc. - Corpus oral : annoté par frontières des phrases et disfluences - Source : chaînes de télévision et stations de radio + 30 min de TuDiCoI
Younes *et al.* (2015)		AL	- Corpus de 31 158 messages – 420 897mots en LDT et 7 145 messages – 160 418 mots en ADT - Lexique L→A 19 763 entrées - Lexique A→L 18 153 entrées - Source : SMS + Chat + forum + Facebook
Graja *et al.* (2015)		A	- Ontologie de 18 concepts - Source : corpus TuDiCol
Harrat *et al.* (2017)	✓	A	- Corpus parallèle multidialectal PADIC : 6 400 phrases pour chaque langue dont 36 648 mots pour le DT (5 dialectes + ASM). - Source : 6 400 phrases du dialecte algérien issues de discours et émissions télévisés algériens enregistrés et transcrits

Tableau 3. *Construction de ressources linguistiques en DT (2) (ADA : avec d'autres dialectes arabes, script : le système d'écriture des ressources (A : arabe / L : latin))*

3.3. *Analyse morphosyntaxique*

L'étude de la littérature montre que plusieurs travaux ont été menés sur l'analyse morphosyntaxique du DT. Dans (McNeil, 2012), un analyseur du DT est présenté. Il permet la segmentation des mots en suivant un ensemble de règles relatives aux suffixes et aux préfixes après une étape de prétraitement consistant à filtrer la ponctuation et les mots étrangers, et à translittérer l'écriture arabe en écriture latine en adaptant la translittération de Buckwalter au DT (Buckwalter, 2004). Zribi *et al.* (2016) se sont, pour leur part, concentrés sur la détection des limites des phrases dans le dialecte tunisien transcrit en utilisant trois approches : une approche fondée sur des règles faites à la main sur la base d'éléments lexicaux et caractéristiques prosodiques ; une approche statistique de classification de mots fondée sur PART (Mohamed *et al.*, 2012) ; une approche hybride combinant les deux approches précédentes. D'autres travaux sur l'analyse morphologique du DT comprennent, notamment, celui de Zribi *et al.* (2013)

Auteurs	ADA	Type de traitement et approche	Évaluation
Belgacem *et al.* (2010)	✓	- Identification de dialectes arabes (9 DA) - GMM	TR : 73,33 %
Neifar *et al.* (2014)		- Système de compréhension de la parole Comp-Dial System (*literal understanding of the Tunisian dialect*) - Adaptation d'un système d'ASM pour DT SARF (*arabic vocal server of railway information*) (Bahou, 2014)	TE : 20,34 %
Lachachi et Adla (2015 ; 2016a ; 2016b)	✓	- Identification de 5 dialectes maghrébins - Modèles de mélange de lois gaussiennes (GMM, UBM-GMM)	Précision : 80,49 %
Hassine *et al.* (2016)	✓	- Reconnaissance de la parole des dialectes maghrébins - FFBPNN et SVM	TR pour FFBPNN : 98,3 % TR pour SVM : 97,5 %
Masmoudi *et al.* (2016)		- Conversion Graphème-en-Phonème - CRF	TE phonétique : 14,09 %. Rappel : 91,41 % Précision : 87,3 %
Masmoudi *et al.* (2017)		- Conversion Graphème-en-Phonème (G2P) et reconnaissance de la parole (RP) - G2P : à base de règles - RP : modèle acoustique de PLP (perceptual linear predictive) et modèle de langage trigramme	G2P : TR niveau phonème : 99,6 % TE niveau mot : 22,6 % RP : TE : 22.6 %
Hassine *et al.* (2018)		- Reconnaissance de la parole du DT - FFBPNN	TR : 98,5 %

Tableau 4. *Traitement de la parole (ADA : avec d'autres dialectes arabes, TR : taux de reconnaissance, TE : taux d'erreur)*

qui a consisté en une adaptation de l'analyseur morphologique existant de l'arabe standard (Al-khalil (Boudlal *et al.*, 2010)) à l'aide d'un ensemble de transformations des patrons verbaux et nominaux de l'ASM en des patrons adaptés au DT, ainsi que la définition d'un ensemble d'affixes et mots-outils du DT. Dans un travail ultérieur, Zribi *et al.* (2017) ont proposé une méthode de désambiguïsation pour l'analyse morphologique du DT et ont testé, pour cela, diverses techniques d'apprentissage automatique (RIPPER (Cohen, 1995), PART (Mohamed *et al.*, 2012) et SVM (Vapnik, 1995)). Ben Moussa Karmani et Alimi (Karmani et Alimi, 2016) ont développé un analyseur morphologique tenant compte des spécificités du DT, et en s'appuyant sur des règles, le WordNet « aebWordNet », un dictionnaire lexical et un système expert linguistique. Par ailleurs, de récents travaux portant sur l'étiquetage grammatical et l'analyse syntaxique du DT commencent à voir le jour. Nous citons en particulier, Boujelbane *et al.* (2014) qui ont eu recours à des ressources existantes pour l'arabe standard, à savoir l'étiqueteur ASM de Stanford entraîné sur une version DT du corpus ATB (Arabic Tree Bank). Cette version a été générée en utilisant un lexique bilingue ASM-DT ainsi qu'un outil de traduction en DT développé par Boujelbane *et al.* (2013a ; 2013b). Dans le même esprit, Hamdi *et al.* (Hamdi *et al.*, 2015) ont également développé un étiqueteur grammatical pour le DT en exploitant sa proximité avec l'ASM. Le processus d'étiquetage est fondé sur des modèles HMM d'ordres différents, entraînés sur 24 k de phrases ASM obtenues du corpus ATB. Pour leur part, Mekki *et al.* (2017) ont travaillé sur la création d'un Treebank arabe tunisien et son utilisation pour effectuer

une analyse syntaxique du DT. Le corpus est la version de la constitution tunisienne écrite en DT qui comprend 12 k mots et 492 phrases. Ils ont utilisé l'analyseur syntaxique de Stanford, principalement dédié à l'ASM, qui reçoit en entrée les phrases tunisiennes normalisées et fournit l'arbre syntaxique de chaque phrase. Ben Moussa Karmani et Alimi. (2016) ont développé un analyseur morphologique tenant compte des spécificités du DT en utilisant une approche fondée sur des règles, le WordNet « aebWordNet », un dictionnaire lexical et un système expert linguistique. Ces divers travaux sont brièvement décrits dans le tableau 5.

Auteur	ADA	Script	Type de traitement et approche	Évaluation
McNeil (2012)		L	- Segmentation des mots - À base de règles	Exactitude : 89,2 %
Zribi *et al.* (2013b)		A	- Analyse morphologique - Adaptation de l'analyseur morphologique Al-Khalil du ASM (Boudlal *et al.*, 2010)	Précision : 80 % F-mesure : 88,86 %
Boujelbane *et al.* (2014a)		A	- Étiquetage grammatical - Adaptation de l'étiqueteur ASM de Stanford	Exactitude : 78,5 %
Hamdi *et al.* (2015)		A	- Étiquetage morphosyntaxique - Adaptation de l'analyseur morphologique MAGEAD (Habash et Rambow, 2006) et étiqueteur HMM	Exactitude : 89 %
Zribi *et al.* (2016)		A	- Segmentation en phrases - À base de règles, statistique, hybride	Précision : 94,8 %
Ben Moussa karmani et Alimi (2016)		A	- Décomposition en morphèmes et analyse morphologique - À base de règles	Précision décomposition : 58,94 % Précision étiquetage : 84,41 %
Zribi *et al.* (2017)		A	- TAMDAS : Système d'étiquetage grammatical - Classification à base de règles PART (Mohamed *et al.*, 2012) et RIPPER (Cohen, 1995), SVM et classifieur bigramme	Exactitude : 87,32 %
Mekki *et al.* (2017)		A	- Analyse syntaxique - Adaptation de l'analyseur syntaxique ASM de Stanford	Précision : 64,43 % F-mesure : 65,58 %

Tableau 5. *Analyse morphosyntaxique (ADA : avec d'autres dialectes arabes, script : le système d'écriture des ressources (A : arabe / L : latin))*

3.4. *Identification de langue sur l'écrit*

L'identification automatique du DT a fait l'objet d'un nombre relativement réduit de travaux. Il s'agit d'identifier le DT dans des écrits pouvant être aussi bien dans le script arabe ou latin. Notons, cependant, que la plupart des travaux recensés ne sont pas vraiment spécifiques à la détection du DT en particulier, mais visent essentiellement à reconnaître l'arabe dialectal en général. Nous avons pu recenser un seul travail sur le dialecte tunisien en particulier, qui est celui de Aridhi *et al.* (2017), sur l'identification des mots du DT transcrits en alphabet latin et dans lequel deux approches ont été expérimentées. La première est fondée sur la méthode N-Gram CSIF (N-Gram Cumulative Sum of Internal Frequencies) de Ahmed *et al.* (2004), et la seconde sur une classification SVM (Support Vector Machines). D'autres travaux ont visé l'identification de l'arabe dialectal en général, y compris le tunisien. Nous citons

principalement ceux de Sadat *et al.* (2014a ; 2014b) qui ont travaillé sur la détection de l'arabe dialectal (couvrant 18 dialectes dont le DT) dans les médias sociaux, en utilisant un classificateur NB (Naive Bayes) fondé sur un modèle de bigrammes de caractères. Le travail de Harrat *et al.* (2015) a porté sur l'identification de 5 dialectes et de l'ASM au niveau des phrases, en utilisant une classification basée sur NB avec les trigrammes de caractères comme caractéristiques. Malmasi *et al.* (2015) ont aussi traité l'identification, au niveau des phrases, de 5 variétés de dialectes avec ASM. Ils ont utilisé le corpus MPCA et ont mené plusieurs expériences avec un classificateur SVM linéaire et un métaclassificateur, en utilisant diverses caractéristiques de surface fondées sur des caractères et des mots. Adouane *et al.* (2016) ont pour leur part, utilisé les SVM pour identifier l'arabe dialectal comprenant plusieurs DA dont le DT. Nous citons enfin, Saadane *et al.* (2017) qui ont comparé une approche linguistique exploitant des dictionnaires avec une approche statistique à base de n-grammes, pour la détection automatique de DA dont le maghrébin (tunisien, algérien et marocain) et l'égyptien. Ces différents travaux sont décrits dans le tableau 6.

Auteur	ADA	Script	Type de traitement et approche	Évaluation
Sadat *et al.* (2014a ; 2014b)	✓	A	- Identification des dialectes arabes (18 DA) - Modèles de langage n-gramme de Markov fondés sur des caractères et classification Naive Bayes	Exactitude : 98 %
Harrat *et al.* (2015)	✓	A	- Identification des dialectes arabes au niveau phrase (5 DA + ASM) - Classification par Naive Bayes basée sur un modèle trigramme de caractères	Précision pour le DT : 68 %
Malmasi *et al.* (2015)	✓	A	- Identification des dialectes arabes au niveau phrase (5 DA + ASM) - Classification SVM et SGM	Exactitude : 74 %
Adouane *et al.* (2016)	✓	A	- Identification des dialectes arabes (8 DA + berbère) - Classification de Cavnar, SVM et PPM fondée sur des modèles n-grammes de caractères et de mots.	F-mesure : 92,94 %
Saadane *et al.* (2017)	✓	AL	- Identification des dialectes arabes au niveau mot (4 DA + ASM) - Approche linguistique à base de dictionnaires + aproche statistique	TE : 14,5 %
Aridhi *et al.* (2017)		L	- Identification du DT au niveau mot - Méthode CSIF + classification SVM	F-mesure - CSIF : 87 % F-mesure - SVM : 90 %

Tableau 6. *Identification (ADA : avec d'autres dialectes arabes, script : le système d'écriture des ressources (A : arabe / L : latin))*

3.5. *Translittération*

La translittération consiste à transformer un mot d'un système d'écriture en un autre tout en préservant sa prononciation. Comme le DT peut être écrit à la fois en arabe et en latin, en particulier sur les réseaux sociaux, les blogs et les forums, des chercheurs se sont intéressés à la tâche de translittération du DT, qui peut être particulièrement utile pour le processus de construction de ressources dialectales et pour de nombreuses autres applications telles que la traduction automatique (traitement des

noms propres), la recherche d'informations, etc. Masmoudi *et al.* (2015) ont abordé la translittération du latin vers l'arabe, en utilisant des règles préétablies selon la norme CODA. Younes *et al.* (2016), se sont orientés vers des méthodes d'apprentissage automatique, en proposant un modèle de translittération fondé sur les HMM. Nous résumons ces deux travaux dans le tableau 7.

Auteur	Sens	Approche adoptée	Évaluation
Masmoudi *et al.* (2015)	L → A	À base de règles	Rappel : de 92 % à 93 %
Younes *et al.* (2016)	L → A	Étiquetage séquentiel (HMM)	Précision : 53 %

Tableau 7. *Translittération (Sens : le sens de la translittération (A : arabe / L : latin))*

3.6. *Traduction*

Les travaux réalisés sur la traduction du DT ont porté essentiellement sur la traduction du DT vers l'ASM, comme ceux réalisés par Hamdi *et al.* (2013a ; 2013b), qui ont exploité des ressources et outils ASM existants pour traduire automatiquement le DT en ASM, ou encore Sadat *et al.* (2014c) qui ont commencé par construire un lexique DT-ASM manuellement comprenant environ 1 600 entrées, ainsi qu'un ensemble de règles de conversion qu'ils ont appliquées à la transformation des verbes du DT vers l'ASM. D'autres travaux ont abordé la traduction, en ASM, de divers dialectes arabes dont le DT tels que Meftouh *et al.* (2015) et Harrat *et al.* (2015) qui ont utilisé le corpus (PADIC), formé d'une collection de 6 400 phrases de dialectes ASM parallèles pour développer un système de traduction, en ayant recours aux outils GIZA++ (Och et Ney, 2003) et SRILM (Stolcke, 2003). L'ensemble de ces travaux sont décrits dans le tableau 8.

Auteur	ADA	Script	Type de traitement et approche	Évaluation
Hamdi *et al.* (2013a ; 2013b)		A	- Traduction de verbes entre DT et ASM - Approche à base de règles	Rappel : de 80 % à 84 %
Sadat *et al.* (2014c)		A	- Traduction DT → ASM de textes - Approche à base de règles	Score BLEU : 14,32
Meftouh *et al.* (2015) Harrat *et al.* (2015)	✓	A	- Traduction entre paires de langues (5 dialectes et ASM) - Techniques de Kneser-Ney et Written-Bell *smoothing*	Score BLEU : 40,48

Tableau 8. *Traduction (ADA : avec d'autres dialectes arabes, script : le système d'écriture des ressources (A : arabe / L : latin))*

3.7. *Analyse de sentiments*

Avec l'utilisation croissante des dialectes arabes dans les médias sociaux, plusieurs travaux ont été initiés sur l'analyse de sentiments (parfois désignée aussi par détection d'opinion), au cours des dernières années. Sayadi *et al.* (2016) ont procédé à la construction de données annotées contenant à la fois du DT et de l'ASM, collectées pendant la période de l'élection de l'Assemblée nationale et de l'élection présiden-

tielle en Tunisie. Ils ont également comparé 5 classificateurs appliqués à la tâche de l'analyse des sentiments (Naive Bayes, SVM, KNN, arbres de décision, forêts d'arbres décisionnels), et mis en œuvre une méthode d'extraction et de sélection des caractéristiques. Ameur *et al.* (2016) se sont focalisés sur l'analyse des sentiments de commentaires en DT, recueillis sur des pages Facebook tunisiennes et ont proposé une méthode pour la construction de dictionnaires émotionnels en distinguant 9 classes émotionnelles (surpris, satisfait, heureux, joyeux, romantique, déçu, triste, en colère et dégoûté). Cette méthode est fondée sur la présence d'émoticônes dans le corpus et sans utiliser de ressources linguistiques externes. Dans le même contexte, Mdhaffar *et al.* (2017) ont commencé par collecter un corpus, appelé TSAC (Tunisian Sentiment Analysis Corpus) à partir de Facebook, qu'ils ont manuellement annoté par les polarités positive et négative. Ils ont utilisé des techniques d'apprentissage automatique (dont le classifieur Perceptron multicouche) pour la classification binaire des commentaires écrits en DT. Le tableau 9 présente un inventaire de ces travaux.

Auteur	ADA	Script	Type de traitement et approche	Évaluation
Sayadi *et al.* (2016)		A	- Construction de ressource pour l'analyse de sentiments (polarité « positive, négative, neutre ») des commentaires issus de Twitter et écrits en ASM et DT. - NB, SVM, KNN, DT, RF	- Exactitude : 71 % - Corpus annoté de 1 754 tweets en DT
Ameur *et al.* (2016)		AL	- Génération des dictionnaires émotionnels en indiquant la polarité : « surpris, satisfait, heureux, joyeux, romantique, déçu, triste, en colère, dégoûté » - Présence d'émoticônes dans les pages Facebook	- F-mesure : 81,01 % - Lexique de 131 937 mots avec polarité
Mdhaffar *et al.* (2017)		AL	- Construction de ressources pour l'analyse de sentiment et détermination de la polarité « positive, négative » des commentaires écrits en DT - SVM, NB et MLP	- TE : 0,22 - Corpus TSAC de 17K commentaires Facebook annotés

Tableau 9. *Analyse de sentiments (ADA : avec d'autres dialectes arabes, script : le système d'écriture des ressources (A : arabe / L : latin))*

3.8. *Codification et normalisation*

La nature informelle des dialectes et leur non-conformité à des règles orthographiques précises rendent difficile leur traitement automatique. Par conséquent, certains chercheurs ont eu recours à une étape intermédiaire susceptible de faciliter leur traitement, à savoir établir des conventions orthographiques. C'est le cas de Zribi *et al.* (2013) qui ont proposé OTTA, une convention orthographique pour la transcription de l'arabe tunisien parlé. Ils ont utilisé les règles orthographiques ASM afin d'en définir de nouvelles qui sont spécifiques au DT. Une autre convention a été développée par Zribi *et al.* (2014) en adaptant le projet CODA (une orthographe conventionnelle pour l'arabe dialectal) de Habash *et al.* (2012) au dialecte tunisien. Le CODA du DT suit les mêmes règles orthographiques que l'ASM avec quelques exceptions et extensions phonologiques, phonolexicales, morphologiques et lexicales. Le système COTA (COnventionalized Tunisian Arabic orthography : orthographe conventionnelle

de l'arabe tunisien) proposé par Boujelbane *et al.* (2016), est un système de normalisation automatique de l'orthographe du DT. Il utilise une méthode hybride combinant des méthodes fondées sur des règles CODA du DT et des méthodes statistiques.

4. Disponibilité des ressources et outils DT

La consultation des plateformes des organisations telles que le Language Data Consortium (LDC) et l'European Language Resources Association (ELRA) permet de constater qu'il existe très peu des ressources pour le traitement du DT, contrairement à l'ASM et à certains autres dialectes arabes (levantin et égyptien). En fait, une simple requête[5] pour les ressources macro-arabes disponibles dans le catalogue LDC donne 164 ressources. Seulement 11 RL multidialectales comprennent le DT. Avec le moteur de recherche ELRA, nous avons trouvé 108 ressources, dont seules 3 RL multidialectales comprennent le DT. Les divers travaux entrepris sur le DT que nous avons présentés dans la section 3 de cet article ont conduit à la construction de diverses RL du DT (de type données telles que des corpus bruts ou annotés, lexiques, dictionnaires, ontologies, etc.). Nous notons cependant que, sur l'ensemble des RL produites dans le cadre de ces travaux, seuls 24 % sont disponibles sur le Web. Le tableau 10 recense ces RL en indiquant si elles sont librement téléchargeables.

RL (auteurs)	Lien	Libre
Corpus TuDiCoI (Graja *et al*., 2010 ; 2013)	https://sites.google.com/site/marwagraja/resources	✓
Corpus TAC (McNeil, 2011 ; 2015)	http://tunisiya.org/ (consultable en ligne)	
Corpus TARIC (Masmoudi *et al.*, 2014a ; 2014b ; 2014c)	http://www-lium.univ-lemans.fr/ bougares/ressources.php https://sites.google.com/site/masmoudiabir/res	✓
Corpus STAC brut et annoté avec les disfluences : audio + transcription (Zribi *et al*., 2015)	https://sites.google.com/site/ineszribi/ressources/corpus	✓
aebWordNet (Ben Moussa Karmani *et al.* 2014, 2015)	https://github.com/nadou12/aebWordNet-Lexicon	✓
Dictionnaire lexical (étiqueté grammaticalement) (Ben Moussa Karmani et Alimi, 2016)	https://github.com/nadou12/Tunisian-Arabic-Lexical-Dictionary	✓
Corpus de 1500 mots (Ben Moussa Karmani et Alimi, 2016)	https://github.com/nadou12/Intelligent-Tunisian-Arabic-Morphological-Analyzer-evaluation-corpus	✓
Corpus TSAC (Mdhaffar *et al*., 2017)	https://github.com/fbougares/tsac	✓
Corpus PADIC (Meftouh *et al.* 2015)	https://sourceforge.net/projects/padic/files/	✓
Lexique DT (Zribi *et al*., 2013b)	https://sites.google.com/site/ineszribi/ressources/lexique	✓
RIO ontology (Karoui *et al.*, 2013a ; 2013b)	https://sites.google.com/site/marwagraja/resources	✓

Tableau 10. *Ressources linguistiques du DT disponibles*

Pour collecter des données afin de construire des lexiques et des corpus, certains chercheurs ont eu recours à différentes ressources librement disponibles sur le Web. Ces ressources sont énumérées dans le tableau 11.

5. Requêtes réalisées le 14 février 2018.

RL (auteurs)	Année	Description - Lien
Documents sélectionnés de la littérature arabe et dialectale (Mohammad Bakri)	2010	Variétés de chansons, théâtres, articles de journaux http://www.langue-arabe.fr/spip.php ?article25
Nouvelle constitution tunisienne (Klibi Salsabil, Hamraoui Salwa, Ben Abda Hana, Gaddes Chawki, Horcheni Farhat, Maalla Anouar)	2014	12 k de mots distribués sur 492 phrases https://www.babnet.net/9/destourderjaaa.pdf
Dictionnaire anglais-tunisien : « *Peace Corps dictionary* » (Rached Ben Abdelkader, Abdeljelil Ayed et Aziza Naouar)	1977	Manuscrit numérisé https://files.eric.ed.gov/fulltext/ed183017.pdf
Dictionnaire tunisien-français « *le Karmous* » (Karim Abdellatif)	2010	3 800 mots, proverbes et expressions https://www.fichier-pdf.fr/2010/08/31/m14401m/dico-karmous.pdf
Dictionnaire tunisien-français	N/A	Plus de 4 k de mots et expressions arabetunisien.com

Tableau 11. *Autres données en DT disponibles*

Nous n'avons pu recenser que deux outils de traitement du DT,(représentant 6 % des travaux présentés dans notre revue) qui sont disponibles sur le Web. Il s'agit de :

– un outil proposé par Zribi *et al.* (2013 ; 2016 ; 2017), qui permet de faire une analyse morphologique, une segmentation en phrases ainsi qu'un étiquetage morphosyntaxique. Cet outil est accessible par le lien : https://sites.google.com/site/ineszribi/ressources/outils-de-traitement-du-dialecte-tunisien ;

– un convertisseur de graphème en phonème (G2P) (Masmoudi *et al.*, 2014b) accessible par le lien : https://sites.google.com/site/masmoudiabir/res.

5. Discussion

5.1. *Portée des travaux de traitement du DT*

Il ressort de cette étude que le traitement automatique du dialecte tunisien suscite, depuis quelques années, un intérêt croissant de la part de plusieurs chercheurs en TAL. Nous avons recensé à ce jour un total de 63 travaux impliquant ce dialecte, dont 50 (79 %) sont dédiés spécifiquement au DT et 13 (21 %) ont porté sur un ensemble de dialectes arabes incluant le DT. Ce nombre reste encore très limité en comparaison de l'arabe standard ou d'autres langues.

En examinant la nature de ces travaux et les types de traitement visés, nous constatons que plus de 43 % d'entre eux ont porté sur la construction de diverses RL pour le DT à cause du manque de ressources indispensables pour travailler sur ce langage. Une part importante de ces travaux (14 %) a été consacrée au traitement de la parole, ce qui s'explique par le fait que le DT constitue une langue essentiellement parlée et très peu écrite. C'est surtout avec la récente croissance de son utilisation sur le Web social que des ressources écrites ont commencé à être collectées et traitées. L'analyse morphosyntaxique ainsi que l'identification des dialectes ont aussi constitué l'objet

de plusieurs travaux (avec respectivement 13 % et 11 % du total des travaux recensés). En revanche, des traitements comme la traduction, l'analyse des sentiments ou encore la translittération restent encore peu abordés pour le DT avec respectivement des pourcentages de 6 %, 5 % et 3 % du total des travaux effectués sur le DT.

Il est également à noter que la plupart des travaux entrepris sur le DT (76 % du total) ont concerné une seule forme écrite de ce dialecte, à savoir celle transcrite dans l'alphabet arabe, malgré la présence importante de contenus dialectaux produits sur le Web en latin. Ainsi, 15 % d'entre eux ont considéré les deux scripts et seulement 9 % d'entre eux ont considéré le DT transcrit en latin. Mis à part le fait que le DT constitue une variante de la langue arabe et est, par conséquent, naturellement écrit en arabe, d'autres raisons peuvent expliquer la focalisation de plusieurs travaux sur la transcription arabe du DT. Parmi ces raisons, l'idée d'exploiter la proximité du DT avec l'ASM afin d'utiliser des ressources et des outils disponibles dédiés à l'arabe pour générer des ressources ou faire des traitements en dialectal, et qui a constitué une approche adoptée par plusieurs chercheurs.

5.2. *Approches de traitement du DT*

Les principales approches adoptées pour le traitement du DT peuvent être classées selon la langue des ressources utilisées ainsi que les méthodes de traitement proposées. Une synthèse de ces approches est présentée dans le tableau 12. Elle distingue les approches fondées sur l'utilisation exclusive de ressources spécifiques au DT de celles utilisant d'autres ressources en ASM et autres dialectes.

Approches	
Ressources utilisées	**Méthode**
DT	Statistique (à base d'apprentissage automatique)
	Linguistique (à base de règles)
	Hybride
ASM	Par adaptation à base de règles, de ressources et d'outils
	Par traduction manuelle de ressources
ASM + autres dialectes arabes	Par traduction manuelle de ressources

Tableau 12. *Synthèse des approches adoptées dans le traitement du DT*

Une comparaison rigoureuse de l'efficacité de ces deux types d'approches ne peut être réalisée, vu le nombre réduit de travaux de part et d'autre pour chaque type d'application, ainsi que l'utilisation de métriques d'évaluation et jeux de données différents. Nous pouvons, cependant, dégager les principales limites et problèmes affectant leur efficacité. Les approches qui s'appuient sur l'utilisation de ressources existantes en ASM et autres dialectes englobent, essentiellement, celles fondées sur la traduction manuelle ou l'adaptation, moyennant des règles préétablies de RL existantes afin de générer leurs équivalents en DT. Elles englobent aussi, les approches fondées sur une

adaptation (essentiellement à base de règles) d'outils d'analyse morphosyntaxique de l'ASM, afin de traiter le DT. Bien que très utiles pour la construction de ressources et d'outils du DT, ces approches posent plusieurs problèmes dont le principal est lié aux étymologies des mots dialectaux. En fait, le DT, comme mentionné dans la section 2, est caractérisé par une interférence linguistique entre l'ASM et d'autres langues comme le français, l'espagnol, le turc, l'italien, etc. Ainsi, de nombreux mots utilisés dans le DT ne proviennent pas de l'ASM et dérivent d'une langue complètement étrangère. De plus, le multilinguisme caractérisant les Tunisiens fait que le DT subit des changements remarquables de jour en jour avec l'introduction de nombreux mots en DT possédant des racines en langues étrangères, auxquelles s'ajoutent de nouveaux suffixes et préfixes et pouvant générer diverses formes fléchies (cf. section 2). Par conséquent, l'idée de s'appuyer uniquement sur l'analogie avec l'ASM serait insuffisante et parfois inefficace. Dans (Almeman et Lee, 2012), on montre que l'analyseur morphologique de l'ASM peut analyser seulement 32 % des mots dialectaux. De plus, l'utilisation de ce type d'approche limite le traitement du DT à celui de sa forme arabe uniquement, ce qui demeure insuffisant pour traiter le dialecte tunisien tel qu'il est largement produit sur le Web. En effet, et comme nous l'avons indiqué dans la section 2.2, l'écriture en latin du dialectal est beaucoup plus utilisée par les Tunisiens dans leurs communications électroniques (SMS, e-mails, Facebook, Twitter, etc.). Les approches ayant recours à des ressources spécifiques au DT se sont principalement fondées sur des méthodes à base de règles. Elles ont été utilisées dans des travaux d'analyse morphosyntaxique du DT, en ayant recours à des ressources annotées de taille relativement réduite (telles que le corpus de McNeil (2012) comprenant uniquement 2 000 mots annotés). Les méthodes fondées sur l'apprentissage automatique ont été principalement adoptées dans des applications telles que l'identification, l'analyse des sentiments et la translittération en utilisant des corpus généralement extraits du Web. Notons cependant, qu'à la date de cette étude et à notre connaissance, des approches à base d'apprentissage profond (*deep learning*), très prisées aujourd'hui dans le domaine du TAL, n'ont pas encore été explorées dans les travaux portant sur le DT. Cela pourrait s'expliquer entre autres par la non-disponibilité de ressources volumineuses indispensables pour l'efficacité de ces méthodes.

5.3. *Ressources linguistiques du DT*

Malgré les efforts fournis dans plusieurs travaux pour la construction de diverses RL pour le DT comme les corpus, lexiques, dictionnaires, etc., celles-ci restent encore relativement limitées en taille (variant de 3 à 800 k mots pour les corpus et de 2 à 40 k entrées pour les lexiques) et en nombre (tableau 11). De plus, ces RL ont souvent concerné le vocabulaire d'un domaine précis et limité. Nous citons à cet effet, les travaux de Graja *et al.* (2010), Karoui *et al.* (Karoui *et al.*, 2013a ; Karoui *et al.*, 2013b) et Neifar *et al.* (2014), dont les données ont concerné les interactions entre le personnel et les clients dans les gares, avec un vocabulaire limité aux réservations, achat de billets, heure, prix, etc., ou encore le travail de Hassine *et al.* (2016) qui a ciblé la prononciation des chiffres de 0 à 9. La construction de ressources de taille et couverture

significatives reste donc une priorité pour pouvoir étudier et traiter le dialecte tunisien, d'autant plus que le Web (et en particulier le Web social) constitue aujourd'hui une source importante de données en dialectal.

6. Conclusion

Cet article présente un état de l'art du traitement automatique du dialecte tunisien. Il commence par une revue des principales caractéristiques linguistiques de ce dialecte ainsi que les difficultés qu'il présente quant à son traitement. Il présente ensuite une revue des principaux travaux ayant porté sur le DT et recense les principales ressources linguistiques et outils TAL actuellement disponibles sur le Web pour ce dialecte. Cette étude a clairement démontré que malgré l'intérêt croissant qu'il a suscité durant ces dernières années, chez de nombreux chercheurs, le traitement du DT est encore à ses débuts et qu'en termes de ressources et outils de TAL qui lui sont dédiés, ce dialecte demeure encore une langue peu dotée. Très peu de travaux ont été effectués sur son analyse linguistique en se limitant au niveau morphosyntaxique. Les applications TAL impliquant le DT se sont principalement limitées à l'identification, la traduction DT-ASM, l'analyse des sentiments et la translittération latin arabe du DT.

Des efforts restent encore à fournir dans la construction de larges ressources linguistiques en DT à rendre disponibles. En cela, les médias sociaux, très largement utilisés par les Tunisiens, constituent une importante source à exploiter. De telles ressources sont indispensables pour l'étude et le développement d'outils et d'applications pour le DT. Elles permettront également de mettre en œuvre des méthodes d'apprentissage profond, qui ont été jusqu'ici très peu explorées dans les travaux sur le DT.

En outre, entreprendre des travaux sur le DT transcrit en latin qui, jusque-là, a été très peu abordé, nous semble d'une grande importance si l'on veut traiter le dialecte tunisien tel qu'il est largement produit sur le Web. Dès lors, les problèmes posés par l'identification du DT et sa translittération vers l'arabe seront des problèmes principaux à résoudre lorsque l'on s'intéresse au traitement des contenus dialectaux produits par les utilisateurs des médias sociaux.

7. Bibliographie

Abdelkader R. B., « Peace Corps English-Tunisian Arabic Dictionary », *ERIC Clearinghouse [Washington, D.C.]*, 1977.

Adouane W., Semmar N., Johansson R., Bobicev V., « Automatic Detection of Arabicized Berber and Arabic Varieties », *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects*, Osaka, Japan, p. 63-72, 2016.

Ahmed B., Cha S.-H., Tappert C., « Language Identification from Text Using N-gram Based Cumulative Frequency Addition », *Proceedings of Student/Faculty Research Day, CSIS, Pace University*, New York, US, 2004.

Almeman K., Lee M. G., « Towards Developing a Multi-Dialect Morphological Analyser for Arabic », *Proceedings of the 4th International conference on Arabic language processing*, Rabat, Morocco, 2012.

Ameur H., Jamoussi S., Hamadou A. B., « Exploiting emoticons to generate emotional dictionaries from facebook pages », *Intelligent Decision Technologies*, p. 39-49, 2016.

Aridhi C., Achour H., Souissi E., Younes J., « Word-Level Identification of Romanized Tunisian Dialect », *Proceedings of the 22nd International Conference on Applications of Natural Language to Information Systems*, Liège, Belgium, p. 170-175, 2017.

Baccouche T., « L'emprunt en arabe moderne », *Beit Al-Hikma–Carthage et I.B.L.V.– Université de Tunis I*, 1994.

Baccouche T., Mejri S., « Atlas linguistique de Tunisie : du littéral au dialectal », *Institut de recherche sur le Maghreb contemporain*, vol. , p. 387-399, 2004.

Bahou Y., « Compréhension Automatique de la Parole Arabe Spontanée : Intégration dans un Serveur Vocal Interactif », *Université de Sfax*, 2014.

Belgacem M., « Construction d'un corpus robuste de différents dialectes arabes », *Proceedings of Actes des VIII èmes RJC Parole*, Avignon, France, 2009.

Belgacem M., Antoniadis G., Besacier L., « Automatic Identification of Arabic Dialects », *Proceedings of the 7th edition of the Language Resources and Evaluation Conference*, Malta, 2010.

Bouamor H., Oflazer N. H. K., « A multidialectal parallel corpus of arabic », *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland, 2014.

Bouchlaghem R., Elkhlifi A., Faiz R., « Tunisian dialect Wordnet creation and enrichment using web resources and other Wordnets », *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, Doha, Qatar, p. 104-113, 2014.

Boudlal A., Lakhouaja A., Azzeddine M., Abdelouafi M., « Alkhalil Morpho Sys1: A Morpho-syntactic analysis System for Arabic texts », *Proceedings of the 2010 International Arab Conference on Information Technology (ACIT'2010)*, Benghazi, Libya, 2010.

Boujelbane R., « Génération de corpus en dialecte tunisien pour l'adaptation de modèles de langage », *Proceedings of Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, 2013.

Boujelbane R., Ellouze M., Béchet F., Belguith L., « De l'arabe standard vers l'arabe dialectal : projection de corpus et ressources linguistiques en vue du traitement automatique de l'oral dans les médias tunisiens », *TAL*, 2014.

Boujelbane R., Khemekhem M. E., Belguith L. H., « Mapping Rules for Building a Tunisian Dialect Lexicon and Generating Corpora », *Proceedings of the International Joint Conference on Natural Language Processing*, Nagoya, Japan, 2013a.

Boujelbane R., Khemekhem M. E., BenAyed S., Belguith L. H., « Building Bilingual Lexicon to Create Dialect Tunisian Corpora and Adapt Language Model », *Proceedings of the 2nd Workshop on Hybrid Approaches to Translation, ACL*, Sofia, Bulgaria, 2013b.

Boujelbane R., Zribi I., et Mariem Ellouze S. K., « An Automatic Process for Tunisian Arabic Orthography Normalization », *Proceedings of the tenth International Conference on Natural Language Processing (HrTAL2016)*, Dubrovnik, Croatia, 2016.

Buckwalter T., « Issues in Arabic orthography and morphology analysis », *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, Geneva, Switzerland, p. 31-34, 2004.

Cohen W. W., « Fast effective rule induction », *Proceedings of the 12th International Conference on Machine Learning*, Tahoe City, California, p. 115-123, 1995.

Eldesouki M., Samih Y., Abdelali A., Attia M., Mubarak H., Darwish K., Kallmeyer L., « rabic Multi-Dialect Segmentation: bi-LSTM-CRF vs. SVM », *3CoRR*, 2017.

Elimam A., « Du Punique au Maghribi Trajectoires d'une langue sémito-méditerranéenne », *Synergies Tunisie*, vol. 1, p. 25-38, 2009.

Elimam A., « Le maghribi, vernaculaire majoritaire à l'épreuve de la minoration », *ENSET – Oran*, 2012.

Elkateb S., Black B., Rodríguez H., Alkhalifa M., Vossen P., Pease A., Fellbaum C., « Building a wordnet for arabic », *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy, 2006.

Graja M., Jaoua M., Belguith L. H., « Lexical Study of A Spoken Dialogue Corpus in Tunisian Dialect », *Proceedings of the International Arab Conference on Information Technology*, Benghazi-Libya, 2010.

Graja M., Jaoua M., Belguith L. H., « Building Ontologies to Understand Spoken Tunisian Dialect », *International Journal of Computer Science, Engineering and Applications (IJCSEA)*, 2011a.

Graja M., Jaoua M., Belguith L. H., « Towards Understanding Spoken Tunisian Dialect », *Proceedings of the 18th International Conference on Neural Information Processing (ICONIP)*, Shanghai, China, 2011b.

Graja M., Jaoua M., Belguith L. H., « Discriminative Framework for Spoken Tunisian Dialect Understanding », *Proceedings of the 1st International Conference on Statistical Language and Speech Processing (SLSP13)*, Tarragona, Spain, 2013.

Graja M., Jaoua M., Belguith L. H., « Statistical Framework with Knowledge Base Integration for Robust Speech Understanding of the Tunisian Dialect », *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2015.

Habash N., Diab M., Rambow O., « Conventional orthography for dialectal Arabic », *Proceedings of the 8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey, 2012.

Habash N., Rambow O., « MAGEAD: A Morphological Analyzer and Generator for the Arabic Dialects », *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, 2006.

Hamdi A., Boujelbane R., Habash N., Nasr A., « The Effects of Factorizing Root and Pattern Mapping in Bidirectional Tunisian - Standard Arabic Machine Translation », *Proceedings of the MT Summit 2013*, Nice, France, 2013a.

Hamdi A., Boujelbane R., Habash N., Nasr A., « Un Système de Traduction de Verbes entre Arabe Standard et Arabe Dialectal par Analyse Morphologique Profonde », *Proceedings of Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, Les Sables d'Olonne, France, 2013b.

Hamdi A., Gala N., Nasr A., « Automatically building a Tunisian Lexicon for Deverbal Nouns », *Proceedings of the 1st Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, Dublin, Ireland, p. 95-102, 2014.

Hamdi A., Nasr A., Habash N., Gala N., « POS-tagging of Tunisian Dialect Using Standard Arabic Resources and Tools », *Proceedings of the Second Workshop on Arabic Natural Language Processing*, Beijing, China, p. 59-68, 2015.

Harrat S., Meftouh K., Abbas M., Jamoussi S., Saad M., Smaili K., « Cross-Dialectal Arabic Processing », *Proceedings of the 16th International Conference on Intelligent Text Processing and Computational Linguistics*, Cairo, Egypt, 2015.

Harrat S., Meftouh K., Smaili K., « Creating Parallel Arabic Dialect Corpus: Pitfalls to Avoid », *Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing*, Budapest, Hungary, 2017.

Hassine M., Boussaid L., Massaoud H., « Tunisian Dialect Recognition Based on Hybrid Techniques », *International Arab Journal of Information Technology*, 2018.

Hassine M., Boussaid L., Messaoud H., « Maghrebian dialect recognition based on support vector machines and neural network classifiers », *International Journal of Speech Technology*, vol. 19, n° 4, p. 687—695, 2016.

Huang J. Z., « Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values », *Data Mining and Knowledge Discovery*, 1998.

Karmani N. B., Alimi A. M., « Construction d'un Wordnet standard pour l'Arabe tunisien », *Colloque pour les Etudiants Chercheurs en Traitement Automatique du Langage naturel et ses applications*, Sousse, Tunisia, 2016.

Karmani N. B., Soussou H., Alimi A. M., « Building a standardized Wordnet in the ISO LMF for aeb language », *Proceedings of the 7th Global Wordnet Conference (GWC 2014), Association for computational linguistics*, Tartu, Estonia, p. 71-77, 2014.

Karmani N. B., Soussou H., Alimi A. M., « Tunisian Arabic aebWordnet: Current state and future extensions », *Proccedings of the 2015 First International Conference on Arabic Computational Linguistics*, Cairo, Egypt, 2015.

Karoui J., Graja M., Boudabous M. M., Belguith L. H., « Domain Ontology Construction from a Tunisian Spoken Dialogue Corpus », *International Conference on Web and Information Technologies (ICWIT 2013)*, Hammamet, Tunisia, 2013a.

Karoui J., Graja M., Boudabous M. M., Belguith L. H., « Semi-automatic Domain Ontology Construction from Spoken Corpus in Tunisian Dialect: Railway Request Information », *International Journal of Recent Contributions from Engineering, Science and IT (iJES)*, vol. 1, n° 1, p. 35-38, 2013b.

Lachachi N., Adla A., « Identification Automatique des Dialectes du Maghreb », *Revue Maghrébine des Langues (RML10)*, vol. , p. 85-101, 2016a.

Lachachi N., Adla A., « Two approaches-based L2-SVMs reduced to MEB problems for dialect identification », *International Journal of Computational Vision and Robotics*, 2016b.

Lachachi N.-E., Adla A., « GMM-Based Maghreb Dialect Identification System », *Journal of Information Processing Systems*, vol. 11, n° 1, p. 22-38, 2015.

Maamouri R., Bies A., Kulick S., Ciul M., Habash N., Eskander R., « Developing an egyptian arabic treebank: Impact of dialectal morphology on annotation and tool development », *Proceedings of the 9th International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, 2014.

Malmasi S., Refaee E., Dras M., « Arabic Dialect Identification using a Parallel Multidialectal Corpus », *Proceedings of the 14th Conference of the Pacific Association for Computational Linguistics (PACLING 2015)*, Bali, Indonesia, p. 209-217, 2015.

Masmoudi A., Bougares F., Ellouze M., Estève Y., Belguith L., « Automatic speech recognition system for Tunisian dialect », *Language Resources and Evaluation*, vol. 52, n° 1, p. 249-267, 2017.

Masmoudi A., Ellouze M., Bougares F., Estève Y., Belguith L., « Conditional Random Fields for the Tunisian Dialect Grapheme-to-Phoneme Conversion », *Proceedings of INTERSPEECH 2016*, San Francisco, USA, 2016.

Masmoudi A., Estève Y., Khmekhem M. E., Bougares F., Belguith L. H., « Phonetic Tool for the Tunisian Arabic », *Proceedings of the 4th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU'14)*, St. Petersburg, Russia, 2014a.

Masmoudi A., Habash N., Ellouze M., Estève Y., Belguith L. H., « Arabic Transliteration of Romanized Tunisian Dialect Text: A Preliminary Investigation », *Proceedings of the 16th International Conference on Intelligent Text Processing and Computational Linguistics*, Cairo, Egypt, 2015.

Masmoudi A., Khemakhem M. E., Estève Y., Belguith L. H., Habash N., « A Corpus and Phonetic Dictionary for Tunisian Arabic Speech Recognition », *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC 2014)*, Reykjavik, Iceland, 2014b.

Masmoudi A., Khemakhem M. E., Estève Y., Bougares F., Dabbar S., Belguith L. H., « Phonétisation automatique du dialecte tunisien », *Proceedings of the 14th Conference of the Pacific Association for Computational Linguistics (PACLING 2015)*, Le Mans, France, 2014c.

McNeil K., « Tunisian Arabic Morphological Parser », *Ling-420*, 2012.

McNeil K., « Tunisian Arabic Corpus: A written corpus of an "unwritten" language », *Proceedings of the International Symposium on Tunisian and Libyan Arabic Dialects*, Vienna, Austria, 2015.

McNeil K., Faiza M., « Tunisian Arabic Corpus: Creating a written corpus of an "unwritten" language », *Proceedings of Workshop on Arabic Corpus Linguistics*, Lancaster, UK, 2011.

Mdhaffar S., Bougares F., Estève Y., Hadrich-Belguith L., « Sentiment Analysis of Tunisian Dialect: Linguistic Resources and Experiments », *Proceedings of the 3rd Arabic Natural Language Processing Workshop (WANLP)*, Valencia, Spain, p. 55-61, 2017.

Meftouh K., Harrat S., Jamoussi S., Abbas M., Smaili K., « Machine Translation Experiments on PADIC: A Parallel Arabic DIalect Corpus », *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, shanghai, China, 2015.

Mejri S., Said M., Sfar I., « Pluringuisme et diglossie en Tunisie », *Synergies Tunisie*, vol. 1, p. 53-74, 2009.

Mekki A., Zribi I., Ellouze M., Belguith L. H., « Syntactic Analysis of the Tunisian Arabic », *Proceedings of the International Workshop on Language Processing and Knowledge Management*, Sfax, Tunisia, 2017.

Mohamed W. N. H. W., Salleh M. N. M., Omar A. H., « A comparative study of reduced error pruning method in decision tree algorithms », *Proceedings of the 2012 IEEE International Conference on Control System, Computing and Engineering*, US, 2012.

Mohand T., « Substrat et convergences: Le berbere et l'arabe nord-africain », *Estudios de Dialectologia*, vol. 4, p. 99-119, 1999.

Mourtada R., Salem F., Alshaer S., « The Arab Social Media Report 2014: Citizen Engagement and Public Services in the Arab World : The Potential of Social Medai », *MBR School of Government*, 2014.

Mzoughi I., « Intégration des emprunts lexicaux au franć cais en arabe dialectal tunisien Linguistique », *Université de Cergy Pontoise*, 2015.

Neifar W., Bahou Y., Graja M., Jaoua M., « Implementation of a Symbolic Method for the Tunisian Dialect Understanding », *Proceedings of the 5th International Conference on Arabic Language Processing (CITALA 2014)*, Oujda, Maroc, 2014.

Novotney S., Schwartz R., Khudanpurn S., « Getting more from automatic transcripts for semi-supervised language modeling », *Computer Speech and Language*, vol. 36, p. 93—109, 2016.

Och F. J., Ney H., « A systematic comparison of various statistical alignment models », *Computational Linguistics archive*, vol. 29, n^{o} 1, p. 19-51, 2003.

Pereira C., « Arabe maghrébin », *Proceedings of Actes du Colloque International Langues d'Europe et de la Méditerranée LEM*, Nice, France, 2005.

Reynolds D. A., Quatieri T. F., Dunn R. B., « Speaker Verification Using Adapted Gaussian Mixture Models », *Digital Signal Processing*, 2000.

Saadane H., Nouvel D., Seffih H., « Une approche linguistique pour la détection des dialectes arabes », *Actes de TALN 2017*, 2017.

Sadat F., Kazemi F., Farzindar A., « Automatic identification of arabic dialects in social media », *Proceedings of the 1st international workshop on Social media retrieval and analysis*, Gold Coast, Australia, p. 35-40, 2014a.

Sadat F., Kazemi F., Farzindar A., « Automatic Identification of Arabic Language Varieties and Dialects in Social Media », *Proceedings of the 2nd Workshop on Natural Language Processing for Social Media (SocialNLP)*, Dublin, Ireland, p. 22-27, 2014b.

Sadat F., Mallek F., Sellami R., Boudabous M. M., Farzindar A., « Collaboratively Constructed Linguistic Resources for Language Variants and their Exploitation in NLP Applications-the case of Tunisian Arabic and the Social Media », *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing*, Dublin, Ireland, p. 102-110, 2014c.

Saidi D., « Développement de la compétence narrative en arabe tunisien : rapport entre formes linguistiques et fonctions discursives », *Université Lyon 2*, 2014.

Salem F., « The Arab Social Media Report 2017: Social Media and the Internet of Things: Towards Data-Driven Policymaking in the Arab World », *MBR School of Government*, 2017.

Sayadi K., Liwicki M., Ingold R., Bui M., « Tunisian Dialect and Modern Standard Arabic Dataset for Sentiment Analysis », *Proceedings of the 2nd International Conference on Arabic Computational Linguistics*, Konya, Turkey, 2016.

Sayahi L., « Diglossia and language contact: Language variation and change in North Africa », *Cambridge University Press*, 2014.

Stolcke A., « SRILM: An Extensible Language Modeling Toolkit », *Proceedings of the 8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH*, Geneva, Switzerland, 2003.

Suwaileh R., Kutlu M., Fathima N., Elsayed T., Lease M., « ArabicWeb16: A New Crawl for Today's Arabic Web », *Proceedings of the 39th annual international ACM SIGIR conference*

on Research and development in information retrieval: SIGIR '16, Pisa, Italy, p. 673-676, 2016.

Vapnik V., « The Nature of Statistical Learning Theory », *Springer New York*, 1995.

Younes J., Achour H., Souissi E., « Constructing Linguistic Resources for the Tunisian Dialect Using Textual User-Generated Contents on the Social Web », *Proceedings of the 1st International Workshop on Natural Language Processing for Informal Text (NLPIT 2015) In conjunction with The International Conference on Web Engineering (ICWE 2015)*, Rotterdam, The Netherlands, 2015.

Younes J., Souissi E., « A quantitative view of Tunisian dialect electronic writing », *Proceedings of the 5th International Conference on Arabic Language Processing*, Oujda, Morocco, p. 63-72, 2014.

Younes J., Souissi E., Achour H., « A Hidden Markov Model for Automatic Transliteration of Romanized Tunisian Dialect », *Proceedings of the 2nd International Conference on Arabic Computational Linguistics*, Konya, Turkey, 2016.

Zbib R., Malchiodi E., Devlin J., Stallard D., Matsoukas S., Schwartz R., Makhoul J., Zaidan O. F., Callison-Burch C., « Machine translation of Arabic dialects », *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montreal, Canada, p. 49-59, 2012.

Zribi I., Boujelbane R., Masmoudi A., Khemekhem M. E., Belguith L. H., Habash N., « A Conventional Orthography for Tunisian Arabic », *Proceedings of the 9th International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, 2014.

Zribi I., Boujelbane R., Masmoudi A., Khemekhem M. E., Belguith L. H., Habash N., « Spoken Tunisian Arabic Corpus STAC: Transcription and Annotation », *Research in Computing Science*, 2015.

Zribi I., Kammoun I., Khemekhem M. E., Belguith L. H., Blache P., « Sentence Boundary Detection for Transcribed Tunisian Arabic », *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, Bochum, Germany, 2016.

Zribi I., Khemekhem M. E., Belguith L. H., Blache P., « Morphological Disambiguation of Tunisian Dialect », *Journal of King Saud University - Computer and Information Sciences*, 2017.

Zribi I., Khmekhem M. E., Belguith L. H., « Morphological analysis of Tunisian Dialect », *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP 2013)*, Nagoya, Japan, 2013.

Notes de lecture

Rubrique préparée par Denis Maurel

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

Elisabetta JEŽEK. The Lexicon: An Introduction (Oxford Textbooks in Linguistics). *Oxford University Press*. 2016. 234 pages. ISBN 978-0-19960-153-0.

Lu par **Kevin BRETONNEL COHEN**

University of Colorado School of Medicine, Computational Bioscience Program, Biomedical Text Mining Group – Université Paris-Saclay

Nous nous trouvons en pleine recrudescence d'une implication de la sémantique dans le domaine du TAL. C'est un essor inouï depuis l'ère des heurts chomskiens-schankiens. Dans un certain sens, cette résurgence de la sémantique dans un contexte de dominance de l'apprentissage machine s'est amorcée avec la naissance de l'étiquetage de rôles sémantiques chez Gildea et Jurafsky, et bat son plein depuis l'avènement du trio PropBank, VerbNet et SemLink. Cet essor s'accélère encore plus vivement avec la renaissance des réseaux neuronaux, l'apprentissage profond et les plongements lexicaux.

La sémantique, de nos jours, se fonde sur une démarche fort différente des précédentes. De la sémantique dite « non profonde » de PropBank à la sémantique proprement dite « profonde » de VerbNet et SemLink, ces prédécesseurs dans notre champ d'enquête se fondent sur des suppositions rationnelles, à savoir sur les socles de la logique, de l'implication et des rôles thématiques. C'est une approche de la représentation du sens qui a comme source un patrimoine millénaire, qui a connu un succès d'une grande portée autant dans les disciplines langagières expérimentales (apprentissage des langues chez les enfants, psycholinguistique…) que dans les disciplines formelles.

Au lieu de la démarche des représentations formelles, les approches fondées sur l'apprentissage profond et les plongements lexicaux sont fortement reliées à une version de la théorie distributionnelle de Firth, ainsi que de celle de Harris. Or, on pourrait dire que la sémantique des plongements lexicaux s'appuie sur une version de la sémantique distributionnelle réduite, voire affaiblie. Tout d'abord, les sémantiques distributionnelles, à l'opposé des plongements lexicaux, ne se restreignent pas aux lexèmes. Loin de cela, la théorie des répartitions lexicales, selon Harris, est explicitement et inéluctablement en lien avec les classes sémantiques, celles-ci étant explicitement définies en tant que telles. La sémantique distributionnelle de Firth s'étend bien au-delà de ces classes sémantiques, allant jusqu'aux contextes sociaux, sémiotiques, etc. On pourrait dire que la sémantique dite « distributionnelle » des plongements lexicaux oublie à la fois ses aïeux, les distributionalistes des années

1950-1970, et ses ancêtres plus lointains, les formalistes. Un rapprochement entre les deux approches est autant possible que désirable. Ce qui nous manque, c'est une ressource pour approfondir la connaissance de la sémantique des composants des plongements lexicaux et autres représentations distributionnelles.

Elisabetta Ježek comble à merveille ce fossé entre la sémantique distributionnelle de nos jours, la sémantique formelle de PropBank, VerbNet et SemLink, la sémantique distributionnelle de Firth et Harris, et les millénaires de théorisation antécédents. Le livre débute avec les composants du sens, soit la problématique de la définition du « mot », l'information lexicale, et le sens lexical. Puis on se tourne vers la structure du lexique, à savoir les structures paradigmatiques et les structures syntagmatiques. Ces dernières créent un pont entre les représentations formelles et les modèles statistiques tels que les collocations. En bref, cet ouvrage traite du lexique dans son sens large, à savoir de sa structure globale, de sa structure paradigmatique, de son lien avec les rapports syntagmatiques et des problématiques théoriques sous-jacentes desdits sujets, ainsi que des approches concrètes vers celles-ci. Dans son contexte plus étendu relatif à d'autres ouvrages dans le domaine de la sémantique lexicale en traitement automatique de langage naturel, ce livre fournit un pivot précieux entre les approches distributionnelles et les approches s'appuyant sur la structure prédicat-argument, ainsi que sur la psycholinguistique, et les ouvrages de sémantique au sens large (à savoir, la sémantique de la phrase, le sens en pragmatique, etc.). Le contenu est présenté très clairement et de façon à ce qu'il soit accessible à tous.

Le premier chapitre introduit les idées de base du lexique, telles que le contraste entre lexique, dictionnaire, et vocabulaire, le lien entre lexique et sémantique, le contraste entre sens lexical et sens grammatical, et surtout la notion du « mot ». Tandis qu'en traitement automatique du langage, on fait de plus en plus la part belle au mot, surtout depuis l'avènement des plongements lexicaux, cette notion du mot, y compris sa circonscription voire son statut ontologique, reste encore contestée, bien qu'elle ait toujours été (et reste encore) sujette à polémique chez les linguistes. On peut dire que ce problème devient de plus en plus important avec l'intérêt accru pour la traduction automatique. En lien avec cette problématique de la circonscription et de la définition du « mot », ce chapitre propose les pistes les plus précieuses et pratiques, tels l'ordre immuable et la cohésion.

Le deuxième chapitre aborde le contenu du lexique. Les données le constituant sont divisées en informations lexicales au sens propre, et informations dites encyclopédiques ou connaissances du monde. La première catégorie exige une distinction entre différents types de sens, notamment la dénotation d'un mot, sa capacité à circonscrire une certaine classe, et ses connotations ou ses associations affectives et de registre. On ajoute à ces deux genres de sens l'acception collocationnelle, le cas échéant. Au-delà de ces genres de signification, l'information lexicale inclut des renseignements sur la forme, que ce soit phonologique ou typographique, et sur la morphologie, cette dernière se trouvant à cheval entre le seuil morphologie et syntaxe. On y rencontre ainsi les parties du discours, la morphologie flexionnelle et dérivationnelle, et concernant les prédicats, le schéma actanciel et, cas exceptionnel, l'Aktionsart.

Ces genres d'informations lexicales ayant été élaborées, ce chapitre s'oriente vers la distinction entre l'information lexicale et la connaissance encyclopédique. Sans prise de position sur cette problématique polémique, ce chapitre expose une gamme de positions : minimaliste, intermédiaire, maximaliste, et inexistante (c'est-à-dire, la position selon laquelle faire une distinction lexicale et encyclopédique n'est pas cohérent).

Le troisième chapitre est consacré au sens. Tandis que les chapitres suivants sont dédiés aux aspects pertinents du lexique. Pour ceux qui s'occupent de la conception de ressources lexicales, ce chapitre est une réflexion bien approfondie sur la définition des acceptions situées au noyau desdites ressources et sur les implications des théories du sens sous-jacentes. De surcroît, c'est dans ce chapitre que l'on apprécie les liens entre le lexique et plusieurs autres sujets de recherche.

D'abord, il traite des caractères des mots qui se révèlent presque toujours les plus problématiques dans le traitement automatique du langage, à savoir, l'ambiguïté et la polysémie. Les cas d'ambiguïtés les plus extrêmes, pourtant les plus faciles à désambiguïser, comme en anglais *pole* (endroit, *magnetic pole*) et *pole* (perche, *fiberglass pole*), ne sont guère traités. En revanche, les cas d'ambiguïtés les plus répandus et (plus) difficiles, à savoir la polysémie des mots possédant plusieurs acceptions étroitement liées, sont bien élaborés.

À partir de là, on se dirige vers plusieurs théories du sens. On peut classifier ces théories comme « internelles » et « relationnelles ».

Parmi les premières, on trouve la théorie appelée référentielle. Selon cette théorie les mots sont utilisés en tant que choses renvoyant à des objets et à des événements qui existent ou qui se passent dans le monde. Cette thèse fait la part belle au monde dans lequel le langage se trouve. Parmi les avantages de cette thèse, on doit compter le fait qu'elle fait partie intégrale de la sémantique dite formelle. Quant aux inconvénients, on constate qu'ils sont difficiles à incorporer dans une approche sur le traitement automatique des langues fréquentiste à cause de son déterminisme.

À cette base référentialiste, l'hypothèse dite « mentaliste » ajoute la notion du « concept », répondant ainsi à une faiblesse de la théorie référentielle, le fait que l'on peut renvoyer à des choses sans existence, par exemple des abstractions non matérielles (*happiness, beauty*) ou non existantes (*unicorns*). À l'opposé du monde extérieur au sein de la référentialité au sens pur, le mentalisme fait la part belle à l'individu. Cette approche vis-à-vis du sens bénéficie d'un lien avec les ontologies qui font partie intégrale des approches computationnelles par rapport à la résolution de la coréférence et à l'informatique médicale fournissant couramment beaucoup de cas d'utilisation du traitement automatique des langues. En revanche, en pratique, elle est limitée par les mêmes problèmes de couverture qui bloquent souvent l'utilisation des ontologies dans des applications du traitement automatique des langues.

Les deux théories précédentes, référentielle et mentaliste, ont en commun des catégories qui sont forcément conscrites. Ceci est un problème pour une conception de la signification, car, de toute évidence, les catégories des concepts n'ont pas de fron-

tières bien définies, mais plutôt des limites assez floues. C'est le cas soit chez les locuteurs d'une langue, soit au sein du lexique, où l'on croise beaucoup d'acceptions connexes subtilement différentes ou souvent difficiles à différencier, même en contexte (*university* en tant qu'organisme, en tant que ses salariés et ses élèves, et en tant que lieu).

La théorie dite *du prototype* émet une conception très différente du concept structuré, précisément une structure avec un noyau – le *prototype* et des composants plus ou moins lointains de ce prototype. Par exemple, le prototype du *concept bird* serait (pour un anglophone) *robin* (rouge-gorge), avec un composant éloigné du noyau *vulture* (vautour) et des composants encore plus lointains *ostrich* (autruche) et *penguin* (manchot). Cette conception plus relationnelle qu'interne de la signification a plusieurs avantages, notamment l'introduction de la similarité à la variabilité des acceptions des mots. Elle apporte aussi l'idée de similarité à la signification, ce qui se révèle primordial pour les approches distributionnelles si courantes de nos jours. On peut donc suivre les traces de la signification depuis les théories relationnelles jusqu'à Aristote, bien qu'il reste à savoir comment concevoir des mises en œuvre informatiques concrètes qui peuvent se servir des mérites des approches aux deux extrémités de cette gamme de possibilités.

Le quatrième chapitre traite du lexique en tant que « structure », à savoir un ensemble organisé de mots connexes. On parle de deux types d'organisation au sein du lexique : le regroupement en classes et les relations du sens. Par « classes » on entend des ensembles de mots regroupés en fonction de comportements morphologiques ou syntaxiques. Un des mérites de ce chapitre est de prendre en considération des idées que l'on rejette parfois facilement, et peut-être trop facilement, en particulier une réflexion approfondie sur les liens entre les parties du discours (*word classes*) et le sens. Un long discours sur la notion de la « catégorie ontologique » de Lyons et sur la « stabilité temporelle » de Givón relie les perspectives linguistiques et morphosyntaxiques aux perspectives philosophiques courantes par le biais des concepts d'endurants, de perdurants, de continuants et d'occurants. Ce discours prolongé se révélera très utile pour tous ceux qui côtoient des ontologues. Ce chapitre prépare la voie qui mène éventuellement au lien avec les représentations distributionnelles du sens, non pas par le biais de la représentation des liens, mais plutôt par celui de la perspective des exigences des mots (telles la structure actancielle (argument structure) et la structure d'événements).

Le cinquième chapitre aborde les structures paradigmatiques au sein du lexique. Cette notion de paradigme comprend les mots qui « rivalisent » entre eux, par exemple synonymes et antonymes, bref, des mots qui peuvent occuper la même position dans une phrase. On croise alors des relations autorisant des implications, telles que l'hyperonymie et l'hyponymie (*vehicle* et, *car*, *move* et *walk*) et la méronymie et l'holonymie (*finger* et *hand*, *athlete* et *team*), ainsi que les complexités de la synonymie et les relations d'opposition. À partir de cela, on se tourne vers les relations de causalité, de finalité, et d'implication (sens *entailment*). Pour les linguistes, ce sont des relations très intéressantes à cause de leur utilité dans l'enquête des sèmes premiers (sémantiques primitives).

Le sixième chapitre est consacré aux relations dites « syntagmatiques ». On aborde ce sujet dans un livre traitant du lexique, d'un ensemble de mots en tant qu'individus, puisque certains mots montrent une répartition limitée par rapport à d'autres mots. Cette répartition étant un caractère d'un mot spécifique, c'est dans le lexique d'une langue que cela est spécifié, bien qu'il s'agisse d'un aspect combinatoire du mot.

Globalement on peut taxonomiser les mots en ce qui concerne leurs contraintes combinatoires en suivant plusieurs axes. Il y a une dimension concernant ce que l'on pourrait qualifier de source des contraintes : des limites probabilistes d'origine conventionnelle, et des limites proprement dites sémantiques ou ontologiques. De surcroît, il y a la présence ou l'absence d'interprétation compositionnelle d'une combinaison de mots. En outre, on peut classer une contrainte par la nature des résultats de sa violation : un énoncé agrammatical ou un énoncé sémantiquement anormal. On peut aussi différencier les combinaisons par la mobilité ou la raideur des relations syntaxiques entre les mots liés. Enfin, on est amené à la conclusion que, contre les affirmations de certains, on ne peut pas considérer les combinaisons de mots (et leurs contraintes) comme un continuum des combinaisons libres aux idiotismes.

Somme toute, on peut conseiller ce livre, agrémenté d'exemples aussi bien nombreux que clairs, à tous ceux qui s'occupent de la conception et de la construction des ressources lexicales et/ou s'intéressent à la sémantique en traitement automatique des langues naturelles ou à la sémantique lexicale, voire générale.

Frédéric LANDRAGIN. Comment parler à un alien? *Le Bélial éditions*. 2018. 263 pages. ISBN 978-2-84344-943-7.

Lu par **Yannis HARALAMBOUS**

IMT Atlantique et UMR CNRS 6285 Lab-STICC

Frédéric Landragin, directeur de recherche au CNRS, membre du laboratoire LATTICE et spécialiste, entre autres, du dialogue homme-machine multimodal, nous offre ici une introduction moderne, claire et pédagogique *à la linguistique, enrichie par sa grande culture et par sa passion pour la science-fiction. Cet ouvrage se caractérise par la générosité de son contenu dans les deux domaines (linguistique et science-fiction), où rien (ou presque) n'a été laissé de côté.*

Une question que l'on peut se poser est celle du public ciblé. Je peux dire avec certitude qu'il y aura au moins deux types de lecteurs : (a) les amateurs de science-fiction éclairés (c'est-à-dire dont la vision de ce genre littéraire ou cinématographique dépasse les sabres laser de Luke Skywalker et le bikini doré de la princesse Léia) qui y trouveront une véritable introduction à la linguistique, introduction qui ne cessera pas de les motiver en leur faisant (re)découvrir des textes de science-fiction emblématiques, et (b) les linguistes en quête d'une introduction à leur discipline accessible au grand public, et plus moderne qu'*Alice au pays du langage* de Marina Yaguello (pour ma part, j'ai fait acheter plusieurs exemplaires de

l'ouvrage de Landragin à la bibliothèque de mon institution puisque je vais m'en servir pour mon cours de TAL). Pourrait-on imaginer un lectorat qui ne s'y connaisse ni en linguistique ni en science-fiction et qui découvrirait ainsi les deux domaines ? Le temps le montrera, et j'imagine que cela fait partie du pari de l'auteur. En tout cas, le choix de la maison d'édition Le Bélial est judicieux : il s'agit d'un « petit » éditeur de science-fiction (« petit » comparé aux éditions du Seuil dont la collection « Folio SF » domine le marché), éditeur qui anime un forum de discussion francophone très actif (870 membres, plus de 60 000 messages) et publie une riche revue trimestrielle (*Bifrost*) depuis plus de vingt ans.

L'ouvrage comporte une introduction et cinq chapitres assez variés. Si le premier chapitre satisfera aussi bien les linguistes que les amis de la science-fiction, le deuxième et le troisième chapitre portent plus sur la linguistique, le quatrième est un excellent mélange de science-fiction et de linguistique, et le cinquième porte plus sur la communication hypothétique avec les aliens, en accord avec le titre de l'ouvrage.

L'introduction, véritable chapitre en soi, fournit les notions fondamentales : on y trouve la différence entre *langue* et *langage* (tout en notant au passage que le titre du roman bien connu *Les langages de Pao* est erroné), les définitions de *signifiant* et *signifié* selon Saussure, la *double articulation*, les *fonctions du langage* selon Jakobson, la *théorie des actes du langage*, ce qu'est et ce que n'est pas la linguistique, etc. Entre deux notions linguistiques, l'auteur réussit à placer des références à des ouvrages ou films de science-fiction : cette alternance thématique permet de mettre en perspective les notions afin de mieux « faire avaler la pilule » du caractère théorique des notions.

Le premier chapitre se focalise sur le sous-genre de la science-fiction qui fait intervenir la linguistique. Au départ, tout en mentionnant un bon nombre d'œuvres connues et moins connues, l'auteur se concentre surtout sur le roman *L'Enchâssement* de Ian Watson, ce qui lui permet d'introduire le côté génératif de la théorie syntaxique de Chomsky. Il enchaîne avec l'hypothèse de Sapir-Whorf, et l'antagonisme entre celle-ci et l'approche chomskyenne, en incluant des exemples d'œuvres de science-fiction dans un camp ou dans l'autre. Il conclut le chapitre avec un bref aperçu de quelques grands classiques de la science-fiction (*Nous autres* de Zamiatine, *1984* d'Orwell, *Babel 17* de Delany, *Légationville* de Miéville) à la lumière des notions étudiées.

Le deuxième chapitre traite de la vision diachronique des langues : leur naissance et leur évolution, leur diversification, leurs familles. On y apprend que le sens des mots évolue, que dans chaque communauté linguistique il y a toujours ceux qui veulent figer une langue et ceux qui la font vivre par un perpétuel changement. Landragin mentionne le mythe raciste du XIX^e^ siècle qui voulait les langues flexionnelles (comme la nôtre, pardi !) plus évoluées que les agglutinantes et les isolantes, mythe utile à rappeler non pas pour son intérêt historique, mais parce qu'un certain nombre d'œuvres de science-fiction s'en servent. Puis, il s'intéresse au futur, avec une section sur l'anticipation linguistique qui permet de distinguer les œuvres sérieuses et réfléchies de science-fiction des produits commerciaux du type

La planète des singes (je parle du film et non pas du livre de notre compatriote Pierre Boulle !) où, deux mille ans après notre ère, une espèce simienne a pris le contrôle de la planète et, comme par hasard, parle... un anglais parfait. Qui dit anticipation linguistique dit anticipation historique et cela permet à Landragin de clore le chapitre par quelques éléments de sociolinguistique, toujours illustrés par des exemples d'œuvres de science-fiction paradigmatiques.

Avant de passer au troisième chapitre, je ne peux m'empêcher de mentionner ce qui constitue, à mon humble avis, le seul « couac » de cet excellent ouvrage : son attachement à la conception phonocentriste. En effet, Landragin affirme que « *la langue est orale et l'écriture n'en est qu'une représentation graphique* » Or, depuis les années 40 des auteurs comme Vachek, Hjelmslev, Uldall, Anis, Catach, Sproat, Coulmas, Dürscheid, Neef et Sébastianoff ont redonné à la modalité écrite de la langue ses lettres de noblesse, et selon eux, celle-ci possède, bel et bien, une double articulation. Il existe une discipline dédiée qui est le pendant de la phonologie : la « graphématique ». Et en Allemagne on parle même de « grapholinguistique » (*Schriftlinguistik*) puisque la modalité écrite touche tous les niveaux d'étude de la langue. Je trouve cela bien réducteur de revenir aux dogmes saussuriens, mais cela bien sûr n'engage que moi.

Dans la lancée du chapitre sur la vie des langues naturelles, le troisième chapitre traite de langues artificielles (thème que l'on retrouve dans l'excellent *Les langues imaginaires* de Marina Yaguello, ainsi que dans *La recherche de la langue parfaite* d'Umberto Eco). Après le passage obligé sur les langues auxiliaires internationales (espéranto et volapük) et une brève section sur les « langues fondées sur la logique » (illustrées par *The Troika Incident* de Brown), l'auteur enchaîne avec les langues fictionnelles, en fournissant un grand nombre d'exemples allant du XIV^{e} siècle jusqu'à la langue dothraki du bien actuel *Game of Thrones*. Mais il ne s'arrête pas là : en effet, il fait la (très subtile) distinction entre langues fictionnelles représentant des langues naturelles et langues fictionnelles représentant des langues artificielles, et il enchaîne avec une autre série d'exemples, tout aussi intéressants, du deuxième cas. Une brève mention de la *glossolalie* (c'est-à-dire l'utilisation d'une soi-disant langue inconnue) permet d'introduire le critère des *hapax* de Marina Yaguello : la distribution des langues naturelles est telle qu'on y trouve entre 46 % et 48 % d'*hapax*, le cas de glossolalie le plus connu (Helene Smith, 1861-1929) ne comportait que 32 % d'hapax, donc il s'agissait très probablement d'une escroquerie.

Suit une section dédiée à la morphologie dérivationnelle, aux néologismes et, plus généralement, aux différents procédés de variation ou de création lexicale adoptés dans la science-fiction. L'auteur donne des conseils aux futurs auteurs de science-fiction sur les pièges à éviter en jouant sur le lexique et la morphologie. La dernière section de ce chapitre est très intéressante puisque l'on entre dans le domaine de l'analyse du discours de la science-fiction, notamment en ce qui concerne la « distanciation cognitive » de Darko Suvin et la « xéno-encyclopédie » d'Irène Langlet, qui caractérisent le processus de lecture de la science-fiction moderne. Cette section donne vraiment envie de se plonger dans les travaux d'analyse des procédés spécifiques à la science-fiction.

Je note au passage que si j'étais l'auteur de ce livre, j'aurais peut-être insisté un peu plus sur la notion de langage contrôlé, sur le fait qu'il existe un véritable continuum entre les langages formels et les langages naturels – et même un indicateur à cinq dimensions pour positionner un langage dans cet espace, donné par Tobias Kuhn –, et sur les tentatives de création de langages de programmation « proches du langage naturel » comme *Inform 7*. D'ailleurs l'auteur ne mentionne qu'une seule fois les langages de programmation dans un paragraphe de dix-sept lignes citant un texte vieux d'une cinquantaine d'années…

Partie centrale de l'ouvrage, le chapitre 4 est particulier et passionnant à lire : l'auteur prend une à une toutes les couches d'étude de la langue (à l'exception, bien sûr, de la graphématique !) et les décrit en s'inspirant d'œuvres de science-fiction. Il commence par le niveau lexical avec quelques notions de terminologie et quelques exemples de structuration de champs lexicaux différant selon les langues (illustrations : les *Borogoves* de Kuttner et Moore, traduit par Boris Vian !, et *LAMA* d'Egan). Puis il passe au niveau phonétique, où il parle évidemment de phonèmes et de l'alphabet phonétique international, mais aussi de langues sifflées. La phonétique est illustrée par l'angoissant *Épépé* de Karinthy où un linguiste se trouve par erreur dans un pays dont il ne comprendra, jusqu'à la dernière page du livre, pas un seul mot de la langue. Une section est dédiée à la prosodie, illustrée par le célèbre « *Je suis désolé, Dave* » énoncé par l'ordinateur HAL (acronyme décalé d'une lettre d'« IBM ») de *2001, Odyssée de l'espace* de Clarke, ordinateur qui fait toujours froid dans le dos aujourd'hui malgré ses cinquante ans bien sonnés.

Suit le niveau morphologique qui, contre toute attente, est fascinant puisqu'on y découvre le *passé antérieur surcomposé de subjonctif futur semi-conditionnel plagal 2e forme* de Douglas Adams, les *préfixes* et *suffixes* de *1984*, les déboires d'un prêtre qui s'est fait mutiler à cause de la mauvaise interprétation d'une dérivation morphologique, les *morphèmes évidentiels* des langues amazoniennes, et le fait que la novlangue de ce même *1984* permet, malgré sa rigoureuse planification, de former des phrases volontairement ambiguës, comme la bien connue « la jeune porte le voile ».

L'exemple qui illustre le niveau syntaxique, tiré de la *Guerre des étoiles*, « *dans le bon ordre les mots placer tu dois* » est immédiatement reconnaissable comme pastiche d'énoncé du maître Yoda (dont la consonance japonaise du nom – qui, en réalité, vient du sanskrit – est sûrement liée au fait que le japonais place le verbe à la fin de la phrase). L'auteur parle d'*ordre figé ou non de mots*, de *cas*, de *langues SVO/SOV/…* et d'*ambiguïtés syntaxiques* comme l'exemple bien connu « j'ai vu l'homme avec un télescope ». La partie qui concerne la syntaxe est peut-être un tantinet moins élaborée que les autres : après une tentative moyennement réussie d'expliquer en quelques lignes (!) la composition sémantique itérative *bottom-up* d'un arbre syntaxique de constituants, l'auteur se rabat sur la *sémantique lexicale*, les figures (*métaphore*, *métonymie*, *synecdoque*) et enfin les *différences de champ lexical* qui rendent difficile la traduction.

Mais, qu'à cela ne tienne, l'auteur se rattrape avec la très riche section consacrée à la pragmatique ! On y apprend que *Babel 17* s'intéresse de près aux *déictiques* et

que dans *Légationville* les phénomènes linguistiques deviennent des référents concrets, des véritables personnages du roman. Ensuite il est question d'*implicatures* (sans les nommer), et enfin, de *performatifs*, illustrés par la nouvelle *L'Histoire de ta vie* de Chiang qui a inspiré le récent film *Premier contact* où les heptapodes connaissent l'avenir et donc tous les actes de leur langage sont performatifs. Puis, on apprend que *Nous autres* de Zamiatine, écrit en Russie en 1920, et ensuite interdit par Staline, décrit une société où les questions sont interdites (idée reprise dans le film *L'Enquête corse*. Enfin, un petit paragraphe sur les *actes locutoires*, *illocutoires* et *perlocutoires* avec un extrait de *La septième fonction du langage* de Binet.

Pour clore ce chapitre, une section sur la stylistique, illustrée par le bel incipit « *J'avais atteint l'âge de mille kilomètres* » du roman *Le Monde inverti* de Priest, qui a la particularité de se situer dans un monde régi par la géométrie hyperbolique où l'avancée du temps correspond à un mouvement physique. Néanmoins, cette section ne parvient pas à montrer clairement ce que Landragin entend par « stylistique ». Pour ma part, je me serais plutôt attendu à y trouver une mention à la nouvelle *A Two-Timer* (en allemand : *Hausfreund von vorgestern* = L'ami d'avant-hier) de David I. Masson, le récit que fait de notre temps (1968) un voyageur temporel, récit rédigé entièrement en anglais du XVII^e^ siècle ! (Notons en passant que David I. Masson, linguiste et accessoirement auteur important de science-fiction, semble avoir totalement échappé à Landragin.) Mais peut-être s'agit-il d'un autre type de stylistique ?

Le cinquième et dernier chapitre de l'ouvrage revient sur son titre : « *Comment parler à un alien ?* » (qui d'ailleurs implique que les aliens sont d'obédience saussurienne et donnent d'office la priorité à la modalité orale de la langue, fait démenti par le film *Premier contact* où l'essentiel de la communication se base sur l'écrit, et pas n'importe quel écrit : un écrit dynamique où la trace écrite évolue dans le temps). Après un récit du décryptage des hiéroglyphes égyptiens et quelques exemples afférents de science-fiction (*Expédition* de Boucher, *Omnilingual* de Piper et, plus près de chez nous et de notre adolescence, *La Nuit des temps* de Barjavel), Landragin nous parle de communication à distance avec les aliens et aboutit tout naturellement aux sondes *Voyager* (qui, conformément au phonocentrisme ambiant au moment de leur lancement, contiennent des centaines d'enregistrements audio de diverses langues, mais quasiment aucun texte écrit), leur fameuse plaque représentant une femme et un homme nus, et le film *Contact* où la délicieuse Jodie Foster discute sur une plage exotique mauve avec un alien ayant pris l'apparence de son père défunt.

Suit une section dédiée à la communication entre les espèces, qui comporte plusieurs exemples d'œuvres de science-fiction répertoriant différents types de communication (y compris, dans *Mémoires d'une femme de l'espace* de Mitchison, celle d'une femme qui communique avec un alien dont le corps tout entier, y compris le sexe, participe à la communication, et en tombe enceinte), ainsi que des exemples où la communication n'a simplement pas abouti (comme dans le grandiose *Solaris* de Lem, porté à l'écran par Tarkovski). On y découvre aussi l'existence d'une discipline scientifique appelée « astrolinguistique » (à ne pas confondre avec l'« astroarchéologie » qui est une pseudoscience s'appuyant sur les théories de von

Däniken), discipline dont le but est d'élaborer une langue destinée à la communication avec les aliens, quels qu'ils soient. Après quelques exemples illustrant la difficulté de cette tâche ainsi que celle du face-à-face hypothétique avec les aliens (discipline appelée « xenolinguistique ») et en introduisant les notions de *deixis* et de *multimodalité*, Landragin conclut le chapitre par un historique des « manuels de langue martienne », allant de 1953 jusqu'aux Utopiales de Nantes et à son ouvrage.

Le livre se termine par un épilogue où l'auteur donne des idées d'extension des propriétés des langues existantes et finit par une exhortation aux nouvelles générations d'auteurs d'imaginer des nouvelles de linguistiques-fictions, pour notre plus grand plaisir. Une bibliographie de douze pages et un index des notions concluent cet ouvrage, que je conseille vivement à tout linguiste, ainsi qu'à tout amateur de science-fiction.

Résumés de thèses

Rubrique préparée par Sylvain Pogodalla

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

Mokhtar Boumedyen BILLAMI : mokhtar.billami@lis-lab.fr

Titre : Désambiguïsation sémantique dans le cadre de la simplification lexicale : contributions à un système d'aide à la lecture pour des enfants dyslexiques et faibles lecteurs

Mots-clés : Désambiguïsation sémantique, simplification lexicale, traitement automatique des langues, enfants dyslexiques, faibles lecteurs.

Title: *Word Sense Disambiguation within Lexical Simplification: Contributions to a Reading Support System for Children with Dyslexia and Poor Readers*

Keywords: *Word sense disambiguation, lexical simplification, natural language processing, dyslexic children, poor readers.*

Thèse de doctorat en Sciences du Langage, Laboratoire d'Informatique et Systèmes (LIS), Département Informatique et Intéractions (DII), Faculté des Sciences, Aix-Marseille Université, sous la direction de Nuria Gala Pavia (MC HDR, Aix-Marseille Université) et Johannes Ziegler (DR, CNRS). Thèse soutenue le 15/11/2018.

Jury : Mme Nuria Gala Pavia (MC HDR, Aix-Marseille Université, codirectrice), M. Johannes Ziegler (DR, CNRS, codirecteur), M. Olivier Ferret (IC HDR, CEA LIST, rapporteur), M. Mathieu Lafourcade (MC HDR, Université Montpellier 2, rapporteur), Mme Cécile Fabre (Pr, Université Toulouse 2, examinatrice), M. Laurent Prévot (Pr, Aix-Marseille Université, examinateur).

Résumé : *La lecture est fondamentale pour tout ce qu'un enfant doit apprendre pendant son parcours scolaire. D'après des rapports nationaux (MJENR 2003) ou internationaux (PISA 2009), 20% à 30% des élèves français sont de faibles lecteurs et ont des difficultés pour comprendre les textes écrits, 5% à 10% sont des enfants dys-*

lexiques. Ces lecteurs sont en grande difficulté face à des textes complexes ou avec un vocabulaire peu courant.

Ces dernières années, un nombre important de technologies ont été créées pour venir en aide aux personnes ayant des difficultés pour lire des textes écrits. Les systèmes proposés intègrent des technologies de la parole (lecture à « voix haute ») ou des aides visuelles (paramétrage ou mise en couleur des polices, ou augmentation de l'espace entre lettres et lignes). Cependant, il est essentiel de proposer aussi des transformations sur le contenu afin d'avoir des substituts de mots plus simples et plus fréquents. Cela permettra de rendre les textes plus accessibles et plus faciles à lire et à comprendre. Le but de cette thèse est de contribuer à un système d'aide à la lecture permettant de proposer automatiquement une version simplifiée d'un texte donné tout en gardant le même sens des mots.

Le travail présenté traite du problème de l'ambiguïté sémantique (très courant en traitement automatique des langues) et vise à proposer des solutions pour la désambiguïsation sémantique à l'aide de méthodes non supervisées et à base de connaissances provenant de ressources lexico-sémantiques. Dans un premier temps, nous proposons un état de l'art sur les méthodes de désambiguïsation sémantique et les mesures de similarité sémantique (essentielles pour la désambiguïsation sémantique). Par la suite, nous comparons divers algorithmes de désambiguïsation sémantique afin d'identifier le meilleur. Enfin, nous présentons nos contributions pour la création d'une ressource lexicale pour le français proposant des synonymes désambiguïsés et gradués en fonction de leur niveau de difficulté de lecture et compréhension. Nous montrons que cette ressource est utile et peut être intégrée dans un module de simplification lexicale de textes.

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-01969248

Chloé CLAVEL : chloe.clavel@telecom-paristech.fr

Titre : Analyse des opinions dans les interactions : de la fouille de données à l'interaction humain-agent

Mots-clés : Informatique affective, traitement automatique des langues, analyse d'opinions, interaction humain-agent, agents conversationnels animés.

Title: *Opinion Analysis in Interactions: from Data Mining to Human-Agent Interaction*

Keywords: *Affective computing, natural language processing, opinion analysis, human-agent interaction, embodied conversational agents.*

Habilitation à diriger des recherches en Informatique, Laboratoire Traitement et Communication de l'Information, UFR 919, Université Pierre et Marie Curie – Paris 6. Habilitation soutenue le 29/05/2017.

Jury : M. Dirk Heylen (Pr, Université de Twente, Pays-Bas, rapporteur), M. Nicolas Sabouret (Pr, Université Paris-Sud, rapporteur), M. Björn Schuller (Pr, Imperial College London, rapporteur), M. Frédéric Béchet (Pr, Aix-Marseille Université, examinateur), M. Patrick Paroubek (Ingénieur de recherche, Université Paris-Sud, examinateur), M. Mohamed Chetouani (Pr, Université Pierre et Marie Curie – Paris 6, président).

Résumé : *Le premier axe de mon activité de recherche concerne l'analyse des opinions dans les interactions humain-humain (conversations téléphoniques des centres d'appels, enquêtes de satisfaction, forums de relation client, données de microblogging données conversationnelles avec le conseiller virtuel d'EDF et interactions face à face avec un agent conversationnel animé). J'ai pu dans ce cadre constituer, à partir des données d'entreprise, des corpus* In-the-wild *riches en expressions spontanées. C'est sur ces corpus et d'autres corpus académiques disponibles que j'ai développé des modèles d'opinions et de sentiments. Les modèles d'opinions construits reposent sur des grammaires constituées de lexiques et de règles linguistiques et sur des méthodes d'apprentissage hybrides permettant d'intégrer des connaissances linguistiques au sein d'algorithmes d'apprentissage automatique. Dans le cadre des corpus issus d'interactions orales, j'ai travaillé sur la caractérisation des phénomènes de parole spontanée (disfluence, parole superposée) et sur la modélisation jointe des paramètres acoustiques et des marqueurs linguistiques. Dans le cadre des corpus issus d'interactions écrites, les structures énonciatives propres à la communication virtuelle et ses spécificités rédactionnelles ont également été étudiées. Dans le contexte de l'interaction humain-agent, mes contributions portent sur la mise en place d'une méthode de détection des opinions et des sentiments qui ancre son fonctionnement dans le contexte de l'interaction, avec la prise en compte, d'une part, du contexte dialogique (paires adjacentes et énoncés précédents de l'utilisateur) et des modalités communicatives de l'agent et, d'autre part, de la structure de Topic donnée par le scénario d'interaction. Par ailleurs, j'ai choisi de travailler sur des modélisations fines du phénomène des opinions/sentiments, en le circonscrivant aux goûts de l'utilisateur (*likes *et* dislikes*) et aux opinions des utilisateurs envers l'interaction (pour la détection d'interactions problématiques).*

Le deuxième axe de mon activité de recherche porte sur les stratégies d'interactions socioémotionnelles de l'agent face à un utilisateur humain et sur la génération chez l'agent d'énoncés porteurs d'attitudes. La complémentarité de ces deux axes repose sur l'intégration du système d'analyse de sentiments décrit dans le paragraphe précédent dans les stratégies d'interactions socioémotionnelles de l'agent. Nous nous sommes focalisés sur le contenu verbal avec la mise en place de stratégies permettant de décider de l'alignement ou non de l'agent sur l'attitude de l'utilisateur et d'instancier les paramètres de réalisation d'un tel alignement. Enfin, concernant la génération des énoncés de l'agent, nous avons travaillé au niveau prosodique, notamment. Deux approches méthodologiques ont été choisies : une approche basée sur l'annotation manuelle des signaux prosodiques suivie d'une analyse statistique des corrélats perceptifs et une approche basée sur l'apprentissage automatique de séquences de si-

gnaux à partir d'un corpus avec l'extraction automatique de descripteurs prosodiques et l'extraction automatique de règles temporelles d'association.

Mnasri MAALI : maali.mnasri@gmail.com

Titre : Résumé automatique multi-document dynamique

Mots-clés : Similarité sémantique, regroupement, ILP, analyse discursive.

Title: *Multi-Document Update-Summarization*

Keywords: *Semantic similarity, clustering, ILP, discourse analysis.*

Thèse de doctorat en Informatique, CEA LIST, LVIC, Université Paris-saclay, sous la direction de Gaël de Chalendar (IC, CEA LIST). Thèse soutenue le 20/09/2018.

Jury : M. Gaël de Chalendar (IC, CEA LIST, directeur), Mme Sophie Rosset (DR, CNRS, LIMSI, présidente), M. Jean-Luc Minel (Pr émérite, Université Paris Nanterre, rapporteur), M. Juan-Manuel Torres-Moreno (MC HDR, Université d'Avignon, rapporteur), M. Antoine Doucet (Pr, Université de La Rochelle, examinateur), M. Olivier Ferret (IC HDR, CEA LIST, encadrant scientifique).

Résumé : *Cette thèse s'intéresse au résumé automatique de texte et plus particulièrement au résumé mis à jour. Cette problématique de recherche vise à produire un résumé différentiel d'un ensemble de nouveaux documents par rapport à un ensemble de documents supposés connus. Elle intègre ainsi dans la problématique du résumé à la fois la question de la dimension temporelle de l'information et celle de l'historique de l'utilisateur. Dans ce contexte, le travail présenté s'inscrit dans les approches par extraction fondées sur une optimisation linéaire en nombres entiers (ILP) et s'articule autour de deux axes principaux : la détection de la redondance des informations sélectionnées et la maximisation de leur saillance. Pour le premier axe, nous nous sommes plus particulièrement intéressés à l'exploitation des similarités interphrastiques pour détecter, par la définition d'une méthode de regroupement sémantique de phrases, les redondances entre les informations des nouveaux documents et celles présentes dans les documents déjà connus. Concernant notre second axe, nous avons étudié l'impact de la prise en compte de la structure discursive des documents, dans le cadre de la Théorie de la Structure Rhétorique (RST), pour favoriser la sélection des informations considérées comme les plus importantes. L'intérêt des méthodes ainsi définies a été démontré dans le cadre d'évaluations menées sur les données des campagnes TAC et DUC. Enfin, l'intégration de ces critères sémantiques et discursifs au travers d'un mécanisme de fusion tardive a permis de montrer dans le même cadre la complémentarité de ces deux axes et le bénéfice de leur combinaison.*

URL où le mémoire peut être téléchargé :
https://tel.archives-ouvertes.fr/tel-01902781
