Traitement automatique des langues

Diversité linguistique
Linguistic Diversity in Natural Language Processing

sous la direction de
Aarne Ranta
Cyril Goutte

Vol. 62 - n°3 / 2021

ATALA

Diversité linguistique
Linguistic Diversity in Natural Language Processing

ATALA
Revue de l'Association pour le Traitement Automatique des Langues

TAL
Vol.
62

n°3
2021

Diversité linguistique
Linguistic Diversity in Natural Language Processing

ATALA

Traitement automatique des langues

Revue publiée depuis 1960 par l'Association pour le Traitement Automatique des LAngues (ATALA), avec le concours du CNRS.

ISSN 1965-0906

https ://www.atala.org/revuetal

Traitement automatique des langues

Traitement automatique des langues

Volume 62 – n°3 / 2021

DIVERSITÉ LINGUISTIQUE
LINGUISTIC DIVERSITY IN NATURAL LANGUAGE PROCESSING

Table des matières

Linguistic Diversity in Natural Language Processing

Aarne Ranta* — **Cyril Goutte****

* *University of Gothenburg, Department of Computer Science and Engineering,* Aarne.Ranta@cse.gu.se
** *Conseil national de recherches Canada, Technologies numériques,* Cyril.Goutte@nrc.ca

ABSTRACT. *Although computational linguistics carries the promise of producing tools for processing and understanding a wide variety of languages, most of the work in NLP still focuses on a small number of languages, and in particular on English. The goal of this special issue is to promote linguistic diversity in NLP, by encouraging the publication of work on languages or language varieties less often studied, as well as methods that can easily and demonstrably be applied to those. Two articles are included in this special issue, one on language identification for building a resource for the Corsican language, the other on machine translation of two indigenous languages of northern Canada.*

RÉSUMÉ. *Bien que la linguistique informatique porte en elle la promesse d'outils aidant au traitement et à la compréhension d'une multitude de langues, la majorité des travaux en TAL porte encore sur un petit nombre de langues, et en particulier sur l'anglais. L'objectif de ce numéro spécial est de promouvoir la diversité linguistique en TAL en encourageant la présentation de travaux portant sur des langues ou variantes de langues moins souvent traitées, ainsi que sur des méthodes qui peuvent être aisément appliquées à celles-ci. Deux articles sont inclus dans ce numéro, l'un sur l'identification de la langue pour constituer une ressource pour la langue corse, l'autre sur la traduction de deux langues autochtones du nord du Canada.*

KEYWORDS: *linguistic diversity, less studied languages.*

MOTS-CLÉS : *diversité linguistique, langues peu traitées.*

TAL. Volume 62 – n°3/2022, pages 7 à 11

1. Introduction

Research in Natural Language Processing (NLP) has largely focused on building various methods, models and tools for handling human language. From its original goal of giving computers the ability to understand and communicate with humans using spoken and written language interaction, it has naturally focused on languages researchers were most familiar with, and in particular on English. The rise of methods based on statistical learning and their reliance on significant amounts of linguistic resources has increased this trend, which the return of neural methods and move to deep learning has further reinforced.

Natural Language Processing systems, especially when they are developed using machine-learning-based techniques, have sometimes been claimed to be language agnostic, suggesting that expanding to more diverse languages may simply be a matter of retraining models on appropriate resources in the target language. However, NLP technology is typically developed on a handful of dominant languages which are sometimes related—when it is not created simply on English. It has been argued (Bender, 2011) that achieving genuine language independence requires a level of linguistic sophistication that is normally not included in NLP systems. A recent study of linguistic diversity (Joshi *et al.*, 2020) suggested a gradation of 6 language groups, from the virtually ignored, to the most dominant. The top two groups, on which most of the NLP work is performed, cover more than 4 billion speakers, but comprise only 25 languages, about 1% of the total number of languages considered (and significantly less than 0.5% of the about 7,000 human languages currently spoken). It is therefore conceivable that many of the linguistic features present in the 99+% of remaining languages are not considered and may pose significant, unforeseen challenges to methods in mainstream NLP. This raises the interesting question of how to complement the mainly computational concern of how methods scale up to more data, with the more pragmatic linguistic concern of how to work on more languages.

As textual resources are critical to feed many of the data-hungry statistical and neural methods, the limited availability of such resources for the vast majority of languages also creates challenges for linguistic diversity in NLP. Low resource NLP was the topic of a recent special issue of this journal (Bernhard and Soria, 2018) and many challenges were described and addressed there: obviously, the lack of resources, but also the heterogeneity both in terms of genre, time or topic, as well as the linguistic heterogeneity due to lack of language normalization or code mixing. These often lead to concerns with the quality of the resources, in addition to their quantity (Caswell *et al.*, 2022).

An additional and significant challenge for linguistic diversity is that it is often difficult to publish work performed on languages other than English. The prevalence of English NLP in the academic literature has long been supported by anecdotal accounts (Munroe, 2015; Mielke, 2016). One can easily speculate over the reasons for this situation. The availability of resources and benchmarks in English probably plays an important role, as it makes it easier to tackle an existing task and show progress using

a proposed new method. This is also due to a clear bias in the perception and assessment of novelty in our field. Novelty is one of the key criterion in many peer-review process, and there is a stronger focus of methodological novelty, while language novelty is typically assessed as "just applying an existing method to a new language".

The goal of this special issue is to favour language diversity in natural language processing by offering a venue for publishing this type of work. We believe this is a timely topic as well. The special theme track for the 2022 conference of the Association for Computational Linguistics is: "Language Diversity: from Low-Resource to Endangered Languages" (ACL, 2022), indicating that the concerns expressed above are shared by the most prominent organization in the field.

2. Summary of the Contributions

This special issue contains two articles addressing very different aspects of language diversity. The first one focuses on resource acquisition and processing tool creation with limited data for that purpose—in that specific case language identification tools. The second paper addresses the issue of building Machine Translation systems, and more specifically the challenges arising from the morphological complexity of polysynthetic languages.

2.1. *L'identification de langue, un outil au service du corse et de l'évaluation des ressources linguistiques*

The first article in this special issue deals with the topic of language identification for Corsican, a language considered endangered by UNESCO. Language identification is a task that is doubly relevant to the topic of this special issue, and offers both challenges and opportunities. First, because although it is a well-known task that has reached near-perfect performance on many languages, it is still challenging in particular when little material is available for training. Secondly, because it is a key language processing tool to filter and identify language-appropriate material in a large collection of documents, in order to build resources for less-studied languages. The paper explores both aspects, adapting and testing a large number of language identifiers on Corsican, and exploring the use of several of these tools to process existing linguistic resources.

2.2. *Towards a Low-Resource Neural Machine Translation for Indigenous Languages in Canada*

The second article is about machine translation for two indigenous languages: Inuktitut and Inuinnaqtun. Inuktitut has almost 40,000 speakers and an official status in Nunavut in Canada, whereas Inuinnaqtun is an endangered language with less than 1,000 native speakers. One obvious challenge of the project is, in particular for

Inuinnaqtun, the scarcity of data. The main focus of the paper is, however, on the polysynthetic nature of both of the languages, requiring a substantial effort in morphological segmentation as preprocessing of machine translation. The outcome is a thorough analysis and evaluation of different methods of segmentation, enabling a neural machine translation system for English-Inuktitut that outperforms the previous state of the art.

Acknowledgements

We wish to thank the editorial committee of the TAL journal for suggesting the topic of this special issue and inviting us to coordinate its scientific committee. We wish to thank more specifically the editors-in-chief for their always patient support during this process and in particular Emmanuel Morin for his invaluable help with SciencesConf. We are indebted to the reviewers and members of the scientific committee who accepted to join us for this special issue and volunteered their time in order to help us select the articles published here: Laurent Besacier (Naver Labs, France), Marine Carpuat (University of Maryland, USA), Leila Kosseim (Concordia University, Canada), Mathieu Mangeot (Université Savoie Mont Blanc, France), Yannick Parmentier (Université de Lorraine, France), Yves Scherrer (University of Helsinki, Finland), and Francis Tyers (Indiana University, USA).

3. References

ACL, "ACL 2022 Theme Track: 'Language Diversity: from Low-Resource to Endangered Languages'", https://www.2022.aclweb.org/post/acl-2022-theme-track-language-diversity-from-low-resource-to-endangered-languages, 2022 (visited February 2022).

Bender E. M., "On Achieving and Evaluating Language-Independence in NLP", *Linguistic Issues in Language Technology*, vol. 6, 2011.

Bernhard D., Soria C., "Traitement automatique des langues peu dotées", *Traitement automatique des langues*, vol. 59, no. 3, 2018.

Caswell I., Kreutzer J., Wang L., Wahab A., van Esch D., Ulzii-Orshikh N., Tapo A. A., Subramani N., Sokolov A., Sikasote C., Setyawan M., Sarin S., Samb S., Sagot B., Rivera C., Gonzales A. R., Papadimitriou I., Osei S., Suarez P. O., Orife I., Ogueji K., Niyongabo R. A., Nguyen T. Q., Muller M., Muller A., Muhammad S. H., Muhammad N. F., Mnyakeni A., Mirzakhalov J., Matangira T., Leong C., Lawson N., Kudugunta S., Jernite Y., Jenny M., Firat O., Dossou B. F. P., Dlamini S., de Silva N., Çabuk Balli S., Biderman S. R., Battisti A., Baruwa A., Bapna A., Baljekar P. N., Azime I. A., Awokoya A., Ataman D., Ahia O., Ahia O., Agrawal S., Adeyemi M., "Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets", *Transactions of the Association for Computational Linguistics*, vol. 10, p. 50-72, 2022.

Joshi P., Santy S., Budhiraja A., Bali K., Choudhury M., "The State and Fate of Linguistic Diversity and Inclusion in the NLP World", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 6282-6293, 2020.

Mielke S. J., “Language diversity in ACL 2004 – 2016”, https://sjmielke.com/acl-language-diversity.htm, December 2016 (visited February 2022).

Munroe R., “Languages at ACL this year”, http://www.junglelightspeed.com/languages-at-acl-this-year/, July 2015 (visited February 2022).

L'identification de langue, un outil au service du corse et de l'évaluation des ressources linguistiques

Laurent Kevers*

* *UMR CNRS 6240 LISA, Università di Corsica - Pasquale Paoli*
Avenue Jean Nicoli, 20250 Corte, France

RÉSUMÉ. *La constitution de corpus est une des premières priorités que rencontrent les langues peu dotées. L'émergence de ressources issues d'Internet, de tailles de plus en plus imposantes et couvrant de nombreuses langues, peut laisser penser que ce point est désormais résolu, ce qui n'est pas le cas. À la suite de Caswell* et al. *(2021), qui ont évalué plusieurs ressources de grande envergure, dont une disposant de contenu corse, nous avons mené une analyse de deux corpus incluant cette langue : An Crúbadán et W2C. Parallèlement à une évaluation manuelle, nous avons estimé la possibilité d'utiliser un ou plusieurs modules d'identification de langue afin de filtrer le contenu de ces ressources, ce qui s'avère possible mais au prix d'un rappel peu élevé. Pour cette tâche, nous avons testé et réentraîné divers systèmes afin de les adapter au mieux au corse. Ce travail nous permet de mettre à disposition un modèle capable d'identifier le corse ainsi que 17 autres langues européennes.*

ABSTRACT. *The constitution of corpora is one of the first priorities faced by less-resourced languages. The emergence of Internet-based resources of increasing size and covering more and more languages may suggest that this issue has been resolved, but this is not the case. Following Caswell* et al. *(2021), who evaluated several large resources, including one with Corsican content, we conducted an analysis of two corpora including this language: An Crúbadán and W2C. In parallel to a manual evaluation, we considered the possibility of using one or more language identification modules to filter the content of these resources, which turns out to be possible but at the cost of low recall. For this task, we tested and re-trained various systems in order to adapt them to Corsican. This work makes it possible to provide a model allowing the identification of 17 European languages as well as Corsican.*

MOTS-CLÉS : *corpus, qualité, identification de langue, langues peu dotées, corse.*

KEYWORDS: *corpora, quality, language identification, less-resourced languages, Corsican.*

TAL. Volume 62 – n° 3/2021, pages 13 à 37

1. Introduction

Dans le domaine des technologies de la langue et du traitement automatique du langage (TAL), une très faible minorité des plus de 7 000 langues répertoriées est dotée de manière satisfaisante de ressources et d'outils. Les langues « numériquement délaissées » font partie, à des degrés divers, de la catégorie des langues « peu dotées ». Selon Joshi *et al.* (2020), qui définissent une nomenclature à six niveaux[1], seules sept langues intègrent le niveau le plus élevé, 65 sont reprises dans les trois niveaux intermédiaires, alors que la vaste majorité des autres langues se situe dans les deux catégories les plus basses !

D'une manière générale, les perspectives pour ces langues sont plutôt pessimistes. Certains estiment qu'à peine 500 langues vivantes pourraient subsister à l'horizon 2100 (Landragin, 2018), alors que d'autres vont même jusqu'à considérer que seules 250 langues pourraient atteindre le statut de « survivant numérique » (Kornai, 2013).

De nombreuses initiatives ont cependant été prises et ont donné lieu à des recommandations techniques et méthodologiques – telles que celles de Berment (2004), Soria *et al.* (2013) ou Ceberio Berger *et al.* (2018) – ainsi qu'à des développements concrets pour diverses langues, dont certaines langues régionales de France (Bernhard *et al.*, 2019 ; Millour, 2020). La situation reste néanmoins préoccupante pour de nombreuses langues.

Parmi les actions à entreprendre pour améliorer le statut de ces langues, la constitution de corpus textuels non bruités, disposant si possible de métadonnées ou d'annotations, et de préférence exploitables au niveau légal, est une priorité. Ces ressources sont importantes car elles permettent de documenter les langues et de mettre au point des outils, en particulier grâce aux techniques d'apprentissage artificiel.

Pour répondre à ce besoin de corpus, il a été de plus en plus courant, depuis environ deux décennies, d'essayer de constituer des ensembles de textes à partir d'Internet. En témoignent diverses manifestations scientifiques, telles que les ateliers Web as Corpus[2], ou plusieurs outils dédiés à cette activité (Baroni et Bernardini, 2004 ; Kilgarriff *et al.*, 2014). Grâce à l'augmentation des capacités de calcul, de transfert et de stockage, d'imposantes collections de documents, telles que ParaCrawl (Esplà-Gomis *et al.*, 2019) ou Common Crawl[3], ont vu le jour. Étant donné leur taille, des outils permettant l'extraction ciblée d'une fraction de leur contenu sont apparus (Roziewski et Stokowiec, 2016 ; Wenzek *et al.*, 2019). Des sous-corpus, qui revendiquent des textes « nettoyés » et organisés par langue – dont certaines sont considérées comme peu dotées – ont également été mis à disposition, entre autres C4Corpus (Habernal *et al.*, 2016), OSCAR (Suárez *et al.*, 2020) ou mC4 (Xue *et al.*, 2021).

La création de corpus pour les langues peu dotées serait-elle un problème résolu ?

1. Allant de « 0 » – très faiblement ou non dotées – à « 5 » – fortement dotées.
2. Voir la page du *Special Interest Group* ACL SIGWAC : `https://www.sigwac.org.uk/`.
3. `https://commoncrawl.org/`

Cela ne semble pas acquis car, à diverses reprises, des doutes ont été émis par rapport à la qualité de ces ressources. Cette impression a été objectivée, en particulier par Caswell *et al.* (2020), Caswell *et al.* (2021) ou Tahir et Mehmood (2021).

D'autre part, le problème des droits d'exploitation est souvent évacué, soit en l'ignorant totalement, soit en le reportant sur l'utilisateur final.

Dans cet article, nous nous plaçons dans le contexte d'une langue peu dotée issue du groupe italo-roman : le corse. Après un point sur sa situation et sur les ressources disponibles dans le contexte des technologies de la langue (section 2), nous nous intéressons à quelques corpus de grande taille qui revendiquent la présence plus ou moins importante de contenu en corse. L'un d'eux a déjà été inspecté par Caswell *et al.* (2021). Comme suggéré et encouragé par les auteurs de cette étude, nous avons examiné deux autres ressources afin d'évaluer si l'identification de langue y est fiable et d'expérimenter une méthode de filtrage visant à ne conserver que le contenu corse. Nous avons choisi de travailler à l'aide de logiciels d'identification de langue, ce qui nous permet, dans le même temps, de faire progresser l'outillage fondamental du corse. Nous nous penchons donc tout d'abord sur ces outils (section 3), avant d'aborder l'évaluation des ressources dans un second temps (section 4).

2. La situation du corse

Le corse fait partie des langues considérées comme étant « en danger » par l'Unesco (Moseley, 2010). L'utilisation du corse au quotidien au travers des applications numériques standard – correction orthographique, traduction automatique, vocalisation, moteurs de recherche, etc. – est de fait très limitée, voire inexistante.

Plusieurs facteurs viennent compliquer l'utilisation du corse dans le domaine numérique. Des variations dialectales sont observées, tant à l'oral qu'à l'écrit. On peut néanmoins identifier cinq aires principales (Dalbera-Stefanaggi, 2002 ; Dalbera-Stefanaggi, 2007), entre lesquelles l'intercompréhension des locuteurs est assurée. Malgré la mise en œuvre d'une approche polynomique (Marcellesi, 1984), l'écriture du corse ne bénéficie pas d'une normalisation. Enfin, le rapport de diglossie qu'entretient le corse avec le français peut se traduire par l'apparition du phénomène d'alternance codique ou d'une tendance à la francisation du lexique, notamment chez les plus jeunes. D'une manière générale, malgré une volonté politique de soutenir la langue, on observe un recul de la pratique du corse au profit du français.

En ce qui concerne les technologies de la langue et le TAL, le corse est versé dans l'avant-dernier groupe de la classification de Joshi *et al.* (2020). Le rapport de l'ELDA de 2014 sur les ressources linguistiques consacrées aux langues de France (Leixa *et al.*, 2014) recense 93 ressources pour le corse, en général de faible ampleur, et dont peu sont disponibles dans un format standard et permettant d'accéder aisément aux données brutes. Ce constat est confirmé par l'inventaire de Kevers *et al.* (2021). Au-delà de quelques corpus récemment créés spécifiquement pour le corse (Kevers et Retali-Medori, 2020), il existe cependant certaines ressources de grande dimension

qui revendiquent l'existence de contenu en corse : An Crúbadán[4] (Scannell, 2007), W2C[5] (Majliš et Zabokrtský, 2012), ou Common Crawl[6] (de 2008 à nos jours).

3. Identification de langue

L'identification de langue, c'est-à-dire l'attribution d'une étiquette représentant la langue d'un texte, est un composant fondamental pour le TAL. À ce titre, il est communément intégré aux chaînes de traitement destinées à gérer plusieurs langues. L'identification correcte de la langue d'un document permet d'appliquer les ressources et méthodes d'analyse les plus appropriées et performantes possible.

Dans le contexte particulier du traitement des langues peu dotées, ce composant revêt toute son importance, car il peut contribuer à la constitution des ressources de base, dont les corpus font partie.

3.1. *Un point sur l'état de l'art*

En raison de l'existence de nombreuses solutions incluant une grande variété de langues, l'identification de langue est une tâche qui est parfois considérée comme résolue. Nous n'allons pas en faire ici un panorama complet, d'autant que d'autres, tels que Jauhiainen *et al.* (2018), s'y sont attelés avant nous.

Les premières approches se sont d'abord intéressées à l'exploitation de listes de mots-clés représentatifs pour chaque langue (Ingle, 1976 ; Giguet, 1995 ; Rehurek et Kolkus, 2009). L'utilisation de n-grammes est ensuite apparue et constitue probablement, avec diverses déclinaisons, le moyen le plus utilisé pour traiter le problème (Dunning, 1994 ; Cavnar et Trenkle, 1994 ; Kerwin, 2006 ; Nakatani, 2010 ; Lui et Baldwin, 2012 ; Majliš, 2012 ; Nakatani, 2012 ; Takçı et Güngör, 2012 ; Brown, 2013). Ces dernières années, des outils basés sur des approches neuronales (Jaech *et al.*, 2016) ou impliquant des plongements lexicaux (Joulin *et al.*, 2017) ont également vu le jour. Quelques-uns de ces systèmes sont exposés plus en détail à la section 3.2.2.

Si ces différentes approches ont en général abouti à des résultats satisfaisants, de nombreux points pour lesquels des progrès doivent encore être apportés ont été mis en évidence. Hughes *et al.* (2006) et Jauhiainen *et al.* (2018), à plus de dix ans d'intervalle, identifient tous deux le support des langues peu dotées, la détection ouverte de langues[7], la prise en compte de documents multilingues, ainsi que les effets des prétraitements, comme n'étant pas encore totalement maîtrisés. D'autres points tels que le support d'un nombre élevé de langues, la distinction entre langues proches et

4. `http://crubadan.org/`
5. `https://ufal.mff.cuni.cz/w2c`
6. `https://commoncrawl.org/`
7. C'est-à-dire la possibilité pour un système de gérer des langues qui lui sont inconnues.

dialectes, ainsi que l'identification de langue pour les textes courts sont également relevés comme problématiques par Jauhiainen *et al.* (2018).

Dans le cadre du traitement automatique du corse, nous sommes confrontés à plusieurs de ces points. Il s'agit d'une langue peu dotée, pour laquelle des variations dialectales sont enregistrées, et qui peut également souffrir d'une certaine proximité linguistique avec l'italien, ainsi que de la situation de diglossie avec le français, celle-ci pouvant se matérialiser par l'alternance de ces deux langues dans certains textes.

Notre objectif prioritaire étant de disposer rapidement d'un module d'identification de langue performant pour le corse, nous avons effectué un inventaire et une évaluation de plusieurs systèmes existants. Certains d'entre eux supportent le corse en standard – nous avons donc pu les utiliser tels quels – d'autres ont dû être adaptés, voire complètement réentraînés.

3.2. *Évaluation et mise au point d'un outil adapté pour le corse*

Au-delà de l'obtention d'un module d'identification de langue le plus précis possible pour le corse, notre travail permet d'aborder plusieurs autres questions. La première concerne la possibilité de mettre au point un système performant de détection de langue au niveau européen en incluant au moins une langue régionale. La seconde porte sur la problématique des textes courts et la capacité des outils à les traiter correctement. Cette question est importante car un outil performant sur de très courtes séquences de caractères pourrait avoir un intérêt dans le traitement de l'alternance codique. Enfin, nous avons effectué la comparaison entre plusieurs solutions entraînées avec des quantités de données plus ou moins élevées, et en imposant, pour certaines évaluations, un équilibre entre les différentes langues.

3.2.1. *Choix des langues cibles*

Le choix des langues à prendre en compte a été conditionné par plusieurs paramètres. Tout d'abord, nous désirions nous placer dans un contexte européen et donc sélectionner en priorité des langues officielles de l'Union européenne. Le second critère est plus d'ordre pratique puisqu'il concerne la disponibilité de données exploitables au niveau technique et légal. Nous avons privilégié le choix d'une source unique pour toutes les langues – la collection collaborative de phrases Tatoeba[8] – afin de travailler sur des textes similaires et de limiter le plus possible tout biais dû à la nature des documents[9]. Nous avons conservé 17 langues sur les 24 officielles de l'UE. Il s'agit de celles qui disposent, dans les données Tatoeba, de plus de 100 000 tokens et de plus de 500 000 caractères. Seuls les textes corses n'ont pas pu être issus de cette source étant donné le très faible nombre de phrases disponibles. Nous avons donc utilisé trois

8. Les données sont publiées sous licence CC BY 2.0 FR sur la page `https://tatoeba.org/fr/downloads`. Le jeu de données a été téléchargé le 24/05/2021.

9. Notre préoccupation étant aussi de ne pas utiliser une ressource trop proche de celles que nous évaluons à la section 4.

corpus disponibles au format XML TEI : A Piazzetta, A Sacra Bìbbia et Wikipedia[10]. Les détails chiffrés relatifs à ces données sont repris au tableau 1.

Langue	**Codes**	**Phrases**	**Tokens**	**Caractères**
Anglais	eng - en	1 473 300	11 260 699	59 491 198
Italien	ita - it	787 115	4 616 453	27 386 085
Allemand	deu - de	549 024	4 332 461	26 982 232
Français	fra - fr	465 299	3 481 188	19 917 382
Portugais	por - pt	385 560	2 697 491	15 215 273
Espagnol	spa - sp	337 010	2 375 025	13 492 307
Hongrois	hun - hu	319 107	1 699 406	11 303 334
Néerlandais	nld - nl	142 681	914 262	5 133 999
Finnois	fin - fi	126 167	632 363	4 676 176
Polonais	pol - pl	109 845	582 059	3 790 406
Lituanien	lit - lt	59 254	282 044	1 862 723
Tchèque	ces - cs	56 177	285 837	1 669 959
Danois	dan - da	49 322	315005	1 723 968
Suédois	swe - sv	41 424	238 152	1 292 812
Grec	ell - el	33 981	180 062	1 069 296
Roumain	ron - ro	24 928	157 743	897 319
Bulgare	bul - bg	24 503	142 862	811 300
Corse	cos - co	-	2 314 619	12 619 463
		Corse - A Piazzetta	516 509	3 001 680
		Corse - A Sacra Bìbbia	867 627	4 144 117
		Corse - Wikipedia	930 493	5 473 666

Tableau 1. *Liste des langues sélectionnées*

3.2.2. *Outils*

Pour la sélection des logiciels, nous avons choisi des outils qui proposent en standard les langues ciblées, ou qui peuvent être réentraînés à partir de nos données. La disponibilité de ces systèmes sous une licence ouverte a également été un point d'attention. Afin de disposer d'une valeur de référence, nous avons codé un système effectuant une identification à partir d'un décompte de mots-clés[11] caractéristiques de

10. Disponibles sous licences CC BY-NC-SA 4.0 et CC BY-SA 3.0 sur `https://bdlc.univ-corse.fr/tal/index.php?page=res`

11. À l'exception de la liste corse, créée pour l'occasion, les mots-clés utilisés sont ceux exploités par Lucene (`https://github.com/apache/lucene`) et proviennent du projet Snowball (`https://github.com/snowballstem`) ou de Jacques Savoy (`http://members.unine.ch/jacques.savoy/clef/`). Le nombre de mots-clés par langue varie entre 78 et 393.

chaque langue. Notre sélection, incluant une estimation du niveau de *Technology readiness level* (TRL[12]), est reprise au tableau 2[13].

Nom	Type	Nb. Lg.	Date	Licence	TRL
Référence	mots-clés	18	-	Dév. personnel	3
Systèmes utilisables sans modification					
YALI	n-grammes	18 (122)	2019	BSD	4
WhatLang	n-grammes	1 475	2015	GPL v.3	4
CLD2	n-grammes	161	2015	Apache v.2	9
FastText	n-grammes	176	2019	MIT, CC BY-SA 3.0	8-9
CLD3	n-grammes	213	2020	Apache v.2	9
Systèmes nécessitant l'ajout du corse					
LibTextCat	n-grammes	18 (163)	2015	BSD	9
Lang. Detect.	n-grammes	18 (53)	2014	Apache v.2	4
Systèmes nécessitant un réentraînement complet					
Langid.py	n-grammes	18 (97)	2017	BSD	4
Ldig	*infinity-gram*	18 (17)	2013	MIT	4
FastText*	n-grammes	18 (176)	2019	MIT	8-9

Tableau 2. *Liste des logiciels sélectionnés*

Nous avons choisi de tester plusieurs logiciels utilisables sans aucune modification. Le premier, YALI[14] (Majliš, 2012), utilise un modèle de langue qui repose sur les listes des cent n-grammes d'octets les plus fréquents pour chaque langue. Chaque n-gramme est accompagné par une probabilité. Le système recherche, parmi les 122 langues supportées, corse inclus, celle obtenant la somme de probabilités la plus élevée en fonction des n-grammes rencontrés. Le logiciel est disponible sous licence BSD.

Avec WhatLang (Brown, 2013), les n-grammes de six, dix ou douze octets sont exploités dans une approche utilisant l'algorithme des k plus proches voisins et une mesure de similarité cosinus. Les n-grammes sont sélectionnés et filtrés selon divers facteurs : leur fréquence, leur taille, leur présence dans des n-grammes plus longs. Des indices négatifs (*stopgrams*) sont aussi utilisés pour certains n-grammes incon-

12. Échelle d'estimation de la maturité d'une technologie allant de 1, « principes de base observés », à 9, « système en production ». Une description complète utilisée par les projets européens Horizon 2020 peut être consultée sur : `https://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/annexes/h2020-wp1415-annex-g-trl_en.pdf`.

13. Le nombre renseigné correspond à la configuration exploitée pour nos tests. Une valeur entre parenthèses indique le nombre total de langues supportées par le système. La date correspond à la dernière modification de celui-ci.

14. `http://ufal.mff.cuni.cz/tools/yali`

nus d'un modèle de langue, mais présent dans d'autres. Enfin, un mécanisme de lissage peut être utilisé afin de favoriser la succession de chaînes de caractères dans une même langue. Initialement capable de reconnaître 1 100 langues, une version élargie à 1 475 langues, corse inclus, a ensuite été proposée[15] sous licence GPL v.3.

CLD2, Compact Language Detector 2[16], est un système bayésien naïf qui exploite des quadrigrammes de caractères. Il s'agit d'un composant de détection de langue développé par Google pour son navigateur Chromium. Il a été initialement présenté comme pouvant reconnaître 83 langues, mais une mise à jour lui permet d'en supporter 78 supplémentaires[17], dont le corse. Ce système est distribué sous licence Apache v.2.

Google a ensuite proposé CLD3[18], une solution neuronale exploitant toujours des n-grammes de caractères. Ce système dispose cette fois du support de 213 langues, incluant le corse, et est proposé sous licence Apache v.2.

Enfin, nous avons repris un dernier logiciel supportant le corse en standard. Il s'agit de FastText[19], une librairie issue du groupe de recherche de Facebook. Elle permet d'apprendre des représentations de données textuelles sous la forme de plongements lexicaux, et de créer des classifieurs de textes (Joulin *et al.*, 2017). Ceux-ci ont été utilisés pour produire un système d'identification de langue prenant en charge 176 langues, dont le corse. La librairie est diffusée sous une licence MIT, alors que les plongements lexicaux sont disponibles sous licence CC BY-SA 3.0.

L'approche TextCat, proposée par Cavnar et Trenkle (1994), est probablement l'une des plus connues. Des n-grammes de caractères de différentes longueurs – n allant de un à cinq – sont exploités pour construire des profils contenant une liste ordonnée des 300 éléments les plus fréquents pour chaque langue. Une mesure de distance, nommée *out-of-place*, consiste à déterminer la somme des distances relevées dans l'ordonnancement des n-grammes d'un document avec celui des différents profils de langue afin d'identifier la langue la plus probable. Diverses implémentations sont disponibles. Nous avons choisi celle intégrée à la suite bureautique Libre Office[20], distribuée sous licence BSD. Le système dispose de profils pour l'ensemble des langues ciblées, sauf pour le corse, pour lequel nous en avons généré un.

Nakatani (2010) propose un autre logiciel permettant la création de modèles de langue, et leur exploitation au travers d'une approche bayésienne naïve. Le logiciel, nommé Language Detection[21], dispose de modèles préentraînés pour 53 langues, mais sans y inclure le corse. L'outil permet cependant la génération de nouveaux modèles, ce que nous avons par conséquent effectué pour le corse. La licence d'utilisation accordée pour cette librairie est Apache v.2.

15. https://sourceforge.net/projects/la-strings/
16. https://github.com/CLD2Owners/cld2
17. https://github.com/CLD2Owners/cld2/wiki/CLD2-Full-Version
18. https://github.com/google/cld3
19. https://fasttext.cc/
20. https://github.com/LibreOffice/libexttextcat
21. https://github.com/shuyo/language-detection/

Enfin, nous avons encore retenu quelques systèmes qui ne supportent pas nativement le corse, mais qui disposent des outils permettant un réentraînement complet d'un modèle. Langid.py[22], proposé par Lui et Baldwin (2012), est le premier de ceux-ci. Ce logiciel utilise des n-grammes d'octets – n étant situé entre un et quatre – afin d'alimenter un classifieur de type bayésien naïf pour lequel une stratégie de sélection de caractéristiques basée sur le gain d'information est mise en place. À l'origine, 97 langues sont supportées, mais la version réentraînée par nos soins a été limitée à nos 18 langues cibles. Langid.py est distribué sous licence BSD.

Le dernier système pris en compte est Ldig[23] (Nakatani, 2012). Celui-ci est conçu pour analyser des documents très courts, tels que des *tweets*. Contrairement à la plupart des autres approches qui extraient des n-grammes de différentes longueurs – par exemple allant de deux à quatre – l'analyse du texte repose ici sur les concepts d'*infinity gram* et de *maximal substring* (Okanohara et Tsujii, 2009). Le principe consiste à énumérer l'ensemble des sous-chaînes de caractères pouvant constituer un document. Étant donné le nombre d'éléments potentiellement très élevé, ceux-ci sont rassemblés en classes d'équivalence, représentées par une sous-chaîne maximale. Ldig, qui est mis à disposition sous licence MIT, dispose du support pour 17 langues. Le modèle généré par nos soins en inclut 18, dont le corse fait bien entendu partie.

Nous avons également décidé de générer un nouveau modèle FastText* à partir de la librairie FastText, qui supporte pourtant officiellement toutes nos langues cibles. Le contraste entre les bons résultats généraux et les très mauvaises performances observées pour le corse nous a cependant incités à lui donner une seconde chance.

3.2.3. *Données*

Comme déjà mentionné, nous exploitons Tatoeba[24], pour les 17 langues européennes, ainsi que les trois corpus corses A Piazzetta, A Sacra Bìbbia et Wikipedia.

Nous avons défini quatre catégories de documents (TSML) : (T) *tiny*, jusqu'à 50 caractères, ce qui pourrait correspondre à une courte alternance de langue dans un texte multilingue ; (S) *small*, à partir de 51 et jusqu'à 300 caractères, soit environ la taille d'un *tweet* ; (M) *medium*, à partir de 301 et jusqu'à 3 000 caractères, taille que l'on pourrait comparer à une page de texte environ ; (L) *large* à partir de 3 001 caractères, catégorie qui représente les documents de plus d'une page. Cette partition permet d'évaluer l'efficacité pour différents contextes ou utilisations.

Le corpus Tatoeba reprend essentiellement des phrases isolées dont la longueur varie, allant jusqu'à plus de 1 500 caractères. A Piazzetta est composé des articles d'un blog journalistique, dont la majorité contient entre 300 et 3 000 caractères. A Sacra Bìbbia, la version corse de la Bible, est organisée selon plusieurs divisions hiérarchiques : parties, chapitres, versets. Des titres peuvent également apparaître à

22. `https://github.com/saffsd/langid.py`
23. `https://github.com/shuyo/ldig`
24. `https://tatoeba.org/fr/downloads`

différents endroits. L'unité de traitement retenue est le chapitre, mais les titres ont été traités séparément. La majorité des « documents » ainsi constitués fait moins de 50 caractères, mais un nombre non négligeable de ceux-ci disposent de 300 à 3 000 caractères, voire plus. Enfin, le corpus Wikipedia est composé en grande partie d'articles faisant entre 50 et 3 000 caractères. Tous ces corpus ont été utilisés pour l'apprentissage et pour les tests. La répartition des corpus selon les catégories TSML est détaillée au tableau 3.

Catégorie	**Tatoeba (17 lg.)**		**A Piazzetta**		**A Sacra Bìbbia**		**Wikipedia**	
T `[1,50]`	4 075 086	81,7 %	79	4,5 %	2 146	64,1 %	41	0,7 %
S `[51,300]`	905 294	18,2 %	223	12,8 %	14	0,4 %	2 760	48,2 %
M `[301,3000]`	4 317	0,1 %	1 179	67,6 %	542	16,2 %	2 532	44,2 %
L `[3001,[`	0	0 %	264	15,1 %	645	19,3 %	395	6,9 %
Total	4 984 697		1 745		3 347		5 728	

Tableau 3. *Répartition TSML des documents en fonction du nombre de caractères*

Le corpus Tatoeba est issu d'une base de données dans laquelle la langue est une métadonnée cruciale. Le corpus A Sacra Bìbbia provient lui d'un ouvrage édité en langue corse. Pour ces données, nous considérons l'attribution d'une langue comme fiable. Les deux autres corpus corses sont issus d'Internet et n'ont pas bénéficié d'une vérification. Ils ne peuvent donc pas prétendre au même niveau de fiabilité. Afin d'écarter les documents qui ne seraient pas majoritairement en corse, nous avons mis en place une étape de filtrage au moyen d'un détecteur de langue par mots-clés élaboré dans le cadre d'un précédent travail. Nous avons ainsi écarté 128 documents du corpus A Piazzetta (7 T, 26 S, 78 M et 17 L), et 141 de Wikipedia (1 T, 38 S, 96 M et 6 L).

Les données présentées ci-dessus ont été réparties en deux ensembles, l'un pour l'apprentissage et l'autre pour le test. Nous n'avons pas de phase de paramétrage qui nécessiterait un ensemble de validation. Afin de minorer les éventuels effets dus à la répartition des données d'apprentissage et de test, nous avons mis en place une validation en cinq plis. Pour les expériences nécessitant un équilibrage des données en fonction de la langue et de la taille des documents, le respect de ces contraintes entraîne une utilisation partielle des données à notre disposition.

Pour chaque pli, un ensemble de test équivalent à environ 150 000 caractères par langue a été réservé, avec une répartition équilibrée entre les différentes catégories TSML. En cas de pénurie pour l'une des catégories, plusieurs documents ont été rassemblés afin de constituer un texte « artificiel » de longueur suffisante. La taille des documents pouvant varier à l'intérieur de chaque catégorie, le nombre de documents n'est pas forcément identique pour chaque langue et pour chaque pli, mais en pratique une certaine régularité a été observée[25]. Pour le corse, en raison de documents

25. La contribution au jeu de test d'une langue va de 136 à 172 documents (153 en moyenne). L'ensemble complet, hors corse, contient entre 2 536 et 2 712 documents (2 605 en moyenne).

en moyenne un peu plus longs, nous avons dû augmenter la taille des ensembles de test à 200 000 caractères afin que le nombre de documents ne chute pas trop[26]. Nous avons également veillé à exploiter les trois corpus de manière équivalente. Pour les ensembles d'apprentissage, deux versions ont été définies. La première (section 3.2.4) est limitée à environ 500 000 caractères par langue[27], répartis équitablement entre les différentes catégories TSML, ce qui nous permet d'obtenir des ensembles d'apprentissage équilibrés en termes de longueur de document et de représentation de chaque langue. Pour la seconde version (section 3.2.5), nous avons décidé d'utiliser la totalité des données à notre disposition, ce qui aboutit à des ensembles déséquilibrés.

En plus de ces données, qui permettent une évaluation sur des documents de même nature et de même source que ceux ayant servi à l'apprentissage, nous avons également mené un test sur des textes complètement différents (section 3.2.6). Nous avons choisi d'exploiter à cet effet le *Digital Corpus of the European Parliament*, DCEP (Hajlaoui *et al.*, 2014), dans sa version nettoyée des balises HTML/XML (*STRIP*)[28]. Il s'agit de documents de différentes natures[29] produits dans le cadre du travail du Parlement européen entre 2001 et 2012. Nous avons sélectionné un sous-corpus ne contenant que les documents disponibles pour l'ensemble de nos 17 langues cibles. À nouveau, nous avons réalisé une répartition des documents en plusieurs catégories en fonction de leur longueur. Nous avons cependant constaté que les documents de moins de 200 caractères s'avéraient très souvent vides de contenu linguistique, ce qui nous a obligés à revoir les catégories précédemment définies. Un autre filtrage a également été mis en place pour écarter les documents ne contenant que des références ou codes, sans inclure véritablement un texte en langue naturelle. Finalement, nous avons sélectionné 3 759 documents disponibles pour chacune des langues[30], répartis de la manière suivante : 198 entre 201 et 500 caractères, 558 entre 501 et 1 000 caractères, 1 882 entre 1 001 et 2 000 caractères, et enfin 1 121 entre 2 001 et 3 000 caractères. Notons enfin que ce corpus ne contient pas de document en corse.

L'ensemble des données textuelles a été soumis à un prétraitement identique. Celui-ci a permis de décapitaliser l'ensemble du texte, de décoller les signes de ponctuation et de normaliser les espacements.

Les résultats des évaluations sont repris aux sections suivantes. Le premier test a pour objectif d'évaluer les différents outils en utilisant des ensembles d'apprentissage équilibrés limités à 500 000 caractères (section 3.2.4). Le deuxième test permet, pour les systèmes identifiés comme les plus performants, d'investiguer les différences obtenues en passant à un ensemble d'entraînement maximal non équilibré (section 3.2.5).

26. Soit 115 documents en moyenne (entre 101 et 129 documents).
27. Cette limite s'est imposée, étant donné la quantité de données disponibles pour la langue la moins dotée dans le cadre de cette évaluation (voir tableau 1).
28. Téléchargeable à l'adresse `https://wt-public.emm4u.eu/Resources/DCEP-2013/DCEP-Download-Page.html`
29. Communiqués de presse, motions, procès-verbaux des sessions plénières, règlement intérieur, rapports, et questions écrites au Parlement.
30. Soit un total de 63 903 documents toutes langues confondues.

Enfin, le troisième test a la vocation de recouper et de vérifier les résultats sur des documents d'une autre nature que ceux ayant servi à l'apprentissage (section 3.2.6).

3.2.4. *Première évaluation : données d'apprentissage limitées mais équilibrées*

L'objectif de cette évaluation est de faire un premier tri parmi l'ensemble des outils que nous avons sélectionnés. Comme exposé précédemment, en plus du système de référence, nous disposons de cinq outils à utiliser sans aucune modification, de deux outils pour lesquels l'ajout du corse a dû être effectué, sans toucher aux autres langues, alors que trois autres outils ont nécessité un réentraînement complet. Pour chacun de ces cinq derniers outils, l'apprentissage de cinq modèles a été effectué à partir d'ensembles de documents d'environ 500 000 caractères chacun. Cinq ensembles de test totalement disjoints ont été utilisés pour l'évaluation. Nous rapportons, au tableau 4, la moyenne des résultats obtenus pour les cinq jeux de données.

Tous les systèmes sont globalement meilleurs que la référence, qui est particulièrement à la traîne pour les documents les plus courts (T et S), ce qui est logique pour une approche par mots-clés. Notons cependant que, pour les documents plus longs, ce système se place dans la même fourchette de valeurs que les autres.

Concernant les 17 langues autres que le corse, les outils pour lesquels nous avons réentraîné un modèle offrent les meilleurs résultats, atteignant une précision de plus de 99 %. Il faut cependant nuancer cette affirmation, puisque la version standard de FastText propose de très bonnes performances. À l'inverse, Langid.py et son modèle complètement réentraîné donnent des résultats assez faibles, s'écartant en moyenne très peu de la référence.

Notons que les systèmes réentraînés ont pu tirer un avantage de la proximité des ensembles d'apprentissage et de test, ce qui est d'ailleurs également le cas de FastText qui intègre Tatoeba dans son ensemble d'apprentissage. D'autre part, les logiciels utilisés de manière standard proposent en général le support d'un nombre plus important de langues, ce qui peut jouer en leur défaveur.

En ce qui concerne le corse, on observe des résultats assez moyens pour les systèmes standard (entre 83,25 % et 90,30 % de précision), voire vraiment catastrophiques pour FastText (précision de 1,24 %). Ce résultat, pouvant provenir d'un fort déséquilibre dans les données d'apprentissage (Siewert *et al.*, 2020), nous a d'ailleurs poussés à effectuer un réentraînement complet de ce système, que nous notons FastText*. Les deux logiciels pour lesquels nous avons effectué un apprentissage uniquement pour le corse s'en sortent mieux, avec une précision d'un peu plus de 92 %. En fin de compte, les meilleures performances sont obtenues avec les systèmes réentraînés complètement pour l'ensemble des langues, à l'exception de Langid.py (83,72 %). Ldig permet de grimper jusqu'à une précision de 94,98 %, alors que FastText* surclasse finalement tous les autres outils en atteignant 97,92 % de précision.

Les meilleurs résultats combinés pour les 18 langues sont logiquement à mettre à l'actif de Ldig (99,12 % de précision) et FastText* (99,42 % de précision).

		Utilisation standard					LEARNcos		LEARNall		
	Réf.	**YALI**	**WhatLang**	**CLD2**	**FastText**	**CLD3**	**TextCat**	**Lang. Detect.**	**Langid.py**	**Ldig**	**FastText***
cos	0,7810	0,9030	0,8678	0,8325	0,0124	0,8820	0,9216	0,9210	0,8372	0,9498	0,9792
eng	0,9469	0,9469	0,9635	0,9800	0,9977	0,9516	0,9447	0,9694	0,8916	0,9882	0,9953
ita	0,8851	0,8709	0,8659	0,8625	0,9940	0,9425	0,8709	0,9186	0,7706	0,9892	0,9904
deu	0,9702	0,9714	0,9845	0,9798	0,9976	0,9833	0,9738	0,9738	0,8452	0,9917	0,9976
fra	0,9267	0,9397	0,9493	0,9597	1,0000	0,9434	0,9268	0,9764	0,8625	0,9847	0,9703
por	0,9002	0,8886	0,8861	0,9436	0,9917	0,9589	0,9130	0,9554	0,8509	0,9800	0,9941
spa	0,9179	0,9369	0,8869	0,9357	0,9952	0,9381	0,9060	0,9583	0,8476	0,9869	0,9976
hun	0,8709	0,9282	0,9661	0,9552	0,9912	0,9409	0,9724	0,9840	0,9327	0,9937	0,9924
nld	0,9736	0,9575	0,8865	0,9300	0,9796	0,9641	0,9861	0,9913	0,9367	0,9988	0,9988
fin	0,8722	0,9894	0,8186	0,9869	0,9935	0,9882	0,9972	1,0000	0,9537	1,0000	1,0000
pol	0,7654	0,9672	0,9673	0,9622	0,9933	0,9755	0,9920	0,9960	0,9636	0,9986	0,9986
lit	0,8899	0,9655	0,9497	0,9773	0,9929	0,9632	0,9672	0,9944	0,9689	0,9986	0,9958
ces	0,8508	0,9472	0,9458	0,9672	0,9816	0,9376	0,9913	0,9942	0,9765	0,9943	0,9971
dan	0,9000	0,9556	0,7591	0,8928	0,9609	0,9083	0,9455	0,9830	0,8894	0,9957	0,9929
swe	0,9474	0,9644	0,9427	0,9373	0,9842	0,9728	0,9674	0,9777	0,9478	0,9944	0,9971
ell	0,8479	1,0000	0,9859	1,0000	1,0000	0,9986	1,0000	1,0000	1,0000	1,0000	1,0000
ron	0,8508	0,9484	0,8513	0,9382	0,9769	0,9554	0,9630	0,9868	0,9594	0,9971	0,9985
bul	0,9709	1,0000	0,9588	0,9182	0,9972	0,9713	1,0000	1,0000	1,0000	1,0000	1,0000
17 lg.	0,8992	0,9516	0,9158	0,9486	0,9899	0,9585	0,9598	0,9800	0,9175	0,9936	0,9951
18 lg.	0,8927	0,9489	0,9132	0,9422	0,9355	0,9542	0,9577	0,9767	0,9130	0,9912	0,9942
T	0,6759	0,8234	0,7236	0,8046	0,9267	0,8381	0,8494	0,9155	0,7293	0,9689	0,9781
S	0,9126	0,9707	0,9438	0,9724	0,9551	0,9841	0,9786	0,9906	0,9119	0,9971	0,9985
M	0,9930	0,9980	0,9888	0,9970	0,9565	0,9970	0,9971	0,9991	0,9977	0,9997	1,0000
L	0,9997	0,9997	0,9982	0,9997	0,9571	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Tableau 4. *Résultats des évaluations pour l'ensemble des outils de détection de langue (apprentissage à 500 k caractères)*

Enfin, l'analyse des résultats obtenus par catégorie de taille de document, nous confirme que la majorité des outils rencontre des difficultés avec les documents courts (T ou S). Les deux meilleurs systèmes mis en évidence ci-dessus – Ldig et FastText* – permettent cependant d'arriver à une très bonne précision pour les documents S (respectivement 99,71 % et 99,85 %), alors que les documents T restent tout de même légèrement en retrait (96,89 % et 97,81 % de précision).

Au vu de ces résultats, Ldig et FastText* nous semblent les outils les plus intéressants, tant au niveau de leurs résultats sur l'ensemble des 17 langues et sur le corse, que de l'efficacité atteinte pour les documents les plus courts.

3.2.5. *Deuxième évaluation : maximisation des données d'apprentissage*

Cette seconde évaluation nous permet d'observer l'effet d'une augmentation des données d'entraînement et de la perte de leur équilibrage. Les ensembles d'apprentissage limités à 500 000 caractères laissent cette fois la place à la totalité des données à notre disposition, telles que décrites au tableau 1. Nous conservons les cinq mêmes ensembles de test que lors de la première évaluation. Ces données n'apparaissent bien entendu pas dans les ensembles d'apprentissage étendus. Cette évaluation a été menée pour les deux meilleurs outils mis en évidence jusqu'ici, Ldig et FastText*. Les résultats détaillés peuvent être consultés au tableau 5, dans lequel sont également repris, à titre de comparaison, les résultats obtenus lors de la première évaluation.

Les différences observées sont plutôt faibles, ce qui est assez logique étant donné les bonnes performances déjà obtenues précédemment. Pour les 17 langues hors corse, Ldig enregistre une progression de 0,54 %, en proposant une précision de 99,90 %. Au contraire, FastText* ne semble pas profiter des données supplémentaires : la précision enregistrée à 99,37 % équivaut à une régression de 0,14 %.

La situation est cependant différente lorsque l'on s'intéresse spécifiquement aux résultats obtenus pour le corse. La précision proposée par Ldig progresse de 1,54 % pour s'établir à 96,52 %, alors que FastText* enregistre, dans ce cas, une amélioration de 1,18 % pour atteindre une précision de 99,11 %.

L'examen des résultats obtenus en fonction des catégories de taille de documents permet de mettre en évidence, pour les deux outils, une amélioration pour la catégorie qui concerne les plus petits documents (T). En revanche, là où Ldig s'améliore pour la catégorie S, et maintient des résultats identiques pour les catégories M et L, on observe, en ce qui concerne FastText*, une légère régression pour les catégories S et M, alors que la catégorie L reste à la précision maximale.

Pour cette évaluation, les évolutions de précision sont contrastées et relativement modestes. Ldig progresse en différents points, alors que FastText* montre à la fois des améliorations et des régressions. Au final, Ldig propose, à 99,71 %, la précision la plus élevée pour les 18 langues (99,36 % pour FastText*), alors que FastText* conserve la meilleure précision pour le corse (99,11 % contre 96,52 % pour Ldig).

	LEARNmin		LEARNmax	
	Ldig	**FastText***	**Ldig**	**FastText***
cos	0,9498	0,9792	0,9652	0,9911
eng	0,9882	0,9953	1,0000	0,9952
ita	0,9892	0,9904	0,9988	0,9856
deu	0,9917	0,9976	1,0000	1,0000
fra	0,9847	0,9703	1,0000	0,9753
por	0,9800	0,9941	0,9988	0,9965
spa	0,9869	0,9976	1,0000	0,9988
hun	0,9937	0,9924	0,9988	1,0000
nld	0,9988	0,9988	0,9988	0,9976
fin	1,0000	1,0000	0,9986	0,9988
pol	0,9986	0,9986	0,9987	0,9972
lit	0,9986	0,9958	0,9986	1,0000
ces	0,9943	0,9971	0,9972	0,9901
dan	0,9957	0,9929	0,9971	0,9873
swe	0,9944	0,9971	0,9985	0,9804
ell	1,0000	1,0000	1,0000	1,0000
ron	0,9971	0,9985	0,9985	0,9914
bul	1,0000	1,0000	1,0000	0,9986
17 lg.	0,9936	0,9951	0,9990	0,9937
18 lg.	0,9912	0,9942	0,9971	0,9936
T	0,9689	0,9781	0,9905	0,9811
S	0,9971	0,9985	1,0000	0,9948
M	0,9997	1,0000	0,9997	0,9983
L	1,0000	1,0000	1,0000	1,0000

Tableau 5. *Résultats des évaluations pour les outils Ldig et FastText*, apprentissage à 500 k caractères (LEARNmin) et avec l'ensemble des données (LEARNmax)*

3.2.6. *Troisième évaluation : corpus DCEP*

Avec cette troisième évaluation, nous voulons vérifier que les résultats obtenus lors des tests précédents ne sont pas (trop) influencés par la proximité des ensembles d'apprentissage et de test utilisés jusqu'ici. Nous avons donc choisi de mettre à l'épreuve les résultats des systèmes qui semblent les plus intéressants jusqu'à présent en les confrontant à un ensemble de test totalement différent et non exploité lors de l'apprentissage. Il s'agit du *Digital Corpus of the European Parliament*, DCEP.

Nous avons sélectionné Ldig et FastText*, entraînés avec les données d'apprentissage maximales. Étant donné l'évolution contrastée de FastText* entre les deux premiers tests, nous avons également choisi d'observer son comportement dans sa version entraînée avec l'ensemble d'apprentissage limité à 500 000 caractères. Bien

que l'ensemble de test soit ici unique – composé par la totalité des données issues de DCEP – les outils choisis disposent chacun de cinq modèles générés lors des étapes précédentes. Les chiffres détaillés présentés au tableau 6 constituent donc à nouveau une moyenne des résultats obtenus avec les cinq modèles disponibles.

	FastText*(LEARNmin)	**Ldig(LEARNmax)**	**FastText* (LEARNmax)**
eng	0,9723	0,9779	0,9305
ita	0,9636	0,9874	0,9043
deu	0,9944	0,9982	0,9941
fra	0,9313	0,9287	0,6764
por	0,9984	0,9984	0,9894
spa	0,9984	0,9985	0,9831
hun	0,9955	0,9979	0,9969
nld	0,9936	0,9968	0,9431
fin	0,9978	0,9987	0,9790
pol	0,9986	0,9950	0,9692
lit	0,9974	0,9961	0,9614
ces	0,9921	0,9980	0,8074
dan	0,9978	0,9974	0,9308
swe	0,9972	0,9979	0,6700
ell	0,9939	0,9983	0,9951
ron	0,9917	0,9936	0,8596
bul	0,9987	0,9892	0,6180
17 lg.	0,9890	0,9910	0,8946
T	0,9945	0,9936	0,9380
S	0,9889	0,9900	0,9233
M	0,9910	0,9921	0,8874
L	0,9847	0,9893	0,8805

Tableau 6. *Résultats des évaluations sur le corpus DCEP*

La précision s'établit pour Ldig à 99,10 %. La majorité des langues bénéficient d'une précision supérieure à 99 %, à l'exception du bulgare (98,92 %), de l'italien (98,74 %), de l'anglais (97,79 %) et du français (92,87 %). La version « minimale » de FastText* suit à peu de choses près la même tendance, avec une précision globale de 98,90 % un peu moins élevée, alors que les langues qui ne franchissent pas le seuil des 99 % sont cette fois limitées à l'anglais (97,23 %), l'italien (96,36 %) ainsi que le français (93,13 %). La version « maximale » de FastText* a donné lieu à de moins bons résultats. La précision globale n'atteint pas les 90 % (89,46 %) et plusieurs langues aboutissent à des résultats très décevants, en particulier pour le bulgare (61,80 %), le suédois (67,00 %) et le français (67,64 %). Nous n'avons pas d'explication objective pour éclairer ces chiffres, mais nous notons cependant que le comportement de cette configuration s'inscrit dans le prolongement de celui observé lors de l'évaluation précédente. Enfin, remarquons que les résultats en retrait pour le français sont observés de manière cohérente pour les trois outils.

À l'issue de ces trois évaluations, nous pouvons constater que Ldig présente des résultats particulièrement intéressants. La précision pour le corse est satisfaisante, et cet outil a pu enregistrer une (légère) progression suite à l'extension des données d'apprentissage. Les résultats se sont également maintenus à un niveau élevé sur des données de nature différente. De plus, une précision supérieure à 99 % a pu être observée pour toutes les catégories de documents, y compris celles représentant les plus courts.

4. Évaluation de ressources

4.1. *Ressources concernées*

Notre objectif est d'obtenir une estimation de la qualité de plusieurs ressources d'assez grande envergure, qui incluent du contenu en corse, et qui ont été constituées à partir d'Internet. La première est proposée par le Crúbadán Project[31] (Scannell, 2007), qui met à disposition des corpus moissonnés pour plus de 2 000 langues. Ces dernières ont été identifiées à l'aide d'une mesure cosinus, et ponctuellement d'un classifieur bayésien naïf, à partir d'un ensemble de trigrammes de référence collectés manuellement. Ce corpus est décliné sous différentes formes : des trigrammes de caractères, des mots simples, ainsi que des bigrammes de mots. Une information de fréquence accompagne chacun des éléments repris dans ces listes. Notre intérêt se porte surtout sur les bigrammes de mots, les autres formes ne se prêtant pas bien à une identification de langue. Cette ressource contient, pour le corse, 541 423 caractères répartis en 50 000 bigrammes, les plus courts n'étant cependant pas toujours pertinents[32].

Le corpus W2C[33] (Majliš et Zabokrtský, 2012) a été constitué à partir de Wikipédia et d'autres sources sur Internet. Il contient plus de 54 Go de données concernant 120 langues, dont l'identification a été réalisée par le système YALI entraîné sur un ensemble initial extrait de Wikipédia. La partie corse, d'une taille de 20 Mo, contient 90 405 lignes représentant chacune un « document ». La taille de ceux-ci varie entre 15 et 188 568 caractères. Le corpus totalise 2 765 040 mots et 16 848 279 caractères.

Enfin, nous nous intéressons aussi aux moissonnages effectués par le projet Common Crawl[34]. La proportion de pages en corse est estimée à 0,24 % de la récolte CC-MAIN-2021-21, ce qui représente 64 146 pages[35]. Les différents moissonnages étant en partie cumulatifs, le nombre total de pages identifiées comme étant exprimées en corse est difficile à déterminer. De plus, étant donné la taille de la ressource, il n'est pas aisé d'accéder au contenu relatif à une langue précise. Divers corpus dérivés de

31. `http://crubadan.org/`
32. À titre d'illustration, le bigramme le plus fréquent est « . \n ».
33. `https://ufal.mff.cuni.cz/w2c`
34. `https://commoncrawl.org/`
35. Ce nombre peut varier d'un moissonnage à l'autre. Les statistiques peuvent être consultées à l'adresse `https://commoncrawl.github.io/cc-crawl-statistics/plots/languages.csv`

Common Crawl ont cependant vu le jour et proposent des sous-ensembles organisés par langues. C'est en particulier le cas du corpus mC4[36] (Xue *et al.*, 2021), qui dispose d'un contenu pour 101 langues, et a été constitué à partir de 71 moissonnages de Common Crawl. Pour l'identification de langue, alors que Common Crawl exploite CLD2, le choix de mC4 s'est porté sur CLD3. La partie corse compte 494 913 extraits, dont on peut estimer qu'ils représentent environ 100 millions de caractères. Le corpus Common Crawl est donc abordé par l'intermédiaire de mC4.

4.2. *Méthodologie*

Les corpus à évaluer ont été soumis à trois outils de détection de langue testés précédemment (section 3.2). Nous avons choisi d'utiliser l'outil de référence, le seul basé sur l'utilisation de listes de mots-clés, le logiciel standard CLD3, ainsi que le système Ldig réentraîné avec le maximum de données pour les 17 langues européennes plus le corse. Ce choix est un compromis entre la diversité des approches ainsi que les performances observées lors de nos tests. CLD3 permet, en outre, de couvrir un nombre élevé de langues, ce qui n'est pas le cas des deux autres outils. Enfin, Ldig nous est apparu comme étant le système proposant le meilleur équilibre entre l'identification du corse et des autres langues, y compris sur des documents très courts.

Il n'est évidemment pas possible de faire une évaluation exhaustive des corpus. Nous avons donc suivi la méthodologie proposée par Caswell *et al.* (2021) et extrait 200 éléments pour chaque corpus. Cet échantillon a été équilibré en sélectionnant des documents en fonction du nombre d'outils les ayant reconnus comme exprimés en corse, soit 25 extraits pour chacune des huit combinaisons possibles. Les données non étiquetées ont été proposées à un expert pour annotation. Cette tâche consiste à attribuer au texte une des catégories suivantes[37] : C pour « (langue) correcte », M pour « (langues) mixées », AL pour « autres langues » et B pour « bruit », c'est-à-dire du contenu non linguistique tel que « ®ªsìálivvàáè nrp¬sõ ». À partir de ces informations et de l'identification automatique de langue, il est possible d'estimer si les différentes combinaisons d'outils peuvent constituer un moyen adéquat pour séparer le bon grain – corse – de l'ivraie – les autres langues et le bruit.

4.3. *Évaluation du corpus mC4*

Ce corpus ayant déjà été évalué par Caswell *et al.* (2021), nous nous faisons ici l'écho de leur travail et reprenons les résultats mis en évidence. Leur audit se base sur cent phrases – issues des 494 913 contenues dans le corpus pour le corse – jugées par un expert. La proportion de phrases effectivement exprimées en corse est de 33 %, dont 2 % sont cependant très courtes, et 2 % sont de faible qualité. Les deux tiers

36. Le jeu de données C4 multilingue est documenté à l'adresse : `https://www.tensorflow.org/datasets/catalog/c4#c4multilingual`

37. Celles-ci s'inspirent des catégories utilisées par Caswell *et al.* (2021).

restants se répartissent pour 48 % en phrases exprimées dans une autre langue, et pour 19 % en éléments non linguistiques. Étant donné ces résultats, et ceux observés pour d'autres langues peu dotées, les auteurs recommandent la plus grande prudence lorsqu'il est fait usage de cette ressource.

4.4. *Évaluation du corpus An Crúbadán*

Nous nous intéressons ici aux bigrammes de mots du corpus An Crúbadán. Un premier filtrage a été effectué afin d'écarter les éléments les plus courts, qui présentent une faible qualité linguistique, soit 25 374 éléments de moins de dix caractères. Les 24 626 bigrammes restants sont caractérisés par une longueur moyenne de 13,42 caractères. Les résultats obtenus suite à l'examen de l'échantillon par l'expert, ainsi que par les trois systèmes d'identification de langue sont présentés au tableau 7.

	C	**M**	**AL**	**B**	**Total**
Jugement expert	184 92,00 %	9 4,50 %	7 3,50 %	0 0,00 %	200
Identification automatique par les trois outils sélectionnés Les échantillons diffèrent par le nombre d'outils ayant reconnu le document comme corse (N). La proportion réelle de chaque combinaison dans le corpus complet est donnée entre parenthèses.					
N = 3 : Ldig + CLD3 + Réf. (8,63 % du corpus)	23 92,00 %	2 8,00 %	0 0,00 %	0 0,00 %	25
N = 2 : Ldig + CLD3 (28,86 % du corpus)	23 92,00 %	2 8,00 %	0 0,00 %	0 0,00 %	25
N = 2 : Ldig + Réf. (2,08 % du corpus)	24 96,00 %	1 4,00 %	0 0,00 %	0 0,00 %	25
N = 2 : CLD3 + Réf. (0,82 % du corpus)	24 96,00 %	1 4,00 %	0 0,00 %	0 0,00 %	25
N = 1 : Ldig (13,50 % du corpus)	19 76,00 %	1 4,00 %	5 20,00 %	0 0,00 %	25
N = 1 : CLD3 (16,0 % du corpus)	25 100,00 %	0 0,00 %	0 0,00 %	0 0,00 %	25
N = 1 : Réf. (0,89 % du corpus)	23 92,00 %	2 8,00 %	0 0,00 %	0 0,00 %	25
N = 0 (29,22 % du corpus)	23 92,00 %	0 0,00 %	2 8,00 %	0 0,00 %	25

Tableau 7. *Évaluation du corpus An Crúbadán*

Les jugements, fournis par notre expert, nous révèlent un niveau de qualité plutôt satisfaisant. En effet, 92 % des éléments examinés ont été jugés comme corrects. À cela, il faut ajouter 4,5 % pour lesquels il y a un mélange de corse et d'autres langues.

Au final, seuls 3,5 % ne constituent pas du contenu corse et ne devraient idéalement pas apparaître dans la ressource.

En ce qui concerne la détection effectuée par les trois outils d'identification de langue, seuls 8,63 % des éléments sont détectés à l'unanimité comme du contenu corse. Au contraire, le cas le plus fréquent est celui où aucun logiciel n'attribue l'étiquette corse (29,22 %), suivi d'assez près par l'identification du corse par le duo Ldig + CLD3 (28,86 %). Vient ensuite la reconnaissance du corse par un seul outil, que ce soit CLD3 (16 %) ou Ldig (13,50 %). Les autres combinaisons, impliquant le système de référence par mots-clés, sont plus marginales.

La confrontation de l'analyse automatique à celle fournie par l'expert ne permet pas de dégager des critères qui offriraient la possibilité d'écarter la totalité du contenu exprimé dans une autre langue sans supprimer une part importante du contenu valide. En effet, ne conserver que les documents ayant été identifiés comme corse par deux outils au minimum éliminerait l'ensemble des AL et un tiers des M, mais ne nous offrirait environ que 40 % du contenu estimé en corse. Une approche plus souple, qui garderait les éléments ayant été identifiés par un outil au minimum, permettrait de faire grimper le rappel à environ 68 %, mais inclurait également quelques éléments exprimés dans d'autres langues. Un filtrage, impliquant une perte de contenu, est donc envisageable sur cette ressource qui, rappelons-le, a été évaluée plutôt positivement par l'expert, mais en ayant tout de même connu une amputation préalable de plus de la moitié de ses données.

4.5. *Évaluation du corpus W2C*

Contrairement au corpus An Crúbadán, W2C propose des textes plus conséquents. Les 200 éléments qui constituent notre échantillon ont une longueur moyenne de 312,31 caractères. Les évaluations réalisées par l'expert et par l'intermédiaire des trois outils d'identification de langue sont reprises au tableau 8.

Le premier élément à mettre en avant concerne la fiabilité de l'identification du corse dans le corpus W2C, qui est nettement en retrait par rapport à ce qui est observé pour An Crúbadán. En effet, à peine 19,50 % des éléments examinés ont été jugés conformes par notre expert. Même si l'on ajoute les 3 % pour lesquels une ou plusieurs autres langues ont été observées, le résultat n'est pas flatteur. La majorité des extraits (67 %) est en réalité exprimée dans une autre langue – l'italien, mais aussi d'autres langues italo-romanes, ainsi que le roumain – alors qu'une part non négligeable de l'échantillon est constituée par du bruit (10,50 %). Même si cette analyse ne porte que sur des données très partielles, les chiffres observés ne laissent que peu de place à l'incertitude quant à la qualité de l'identification du corse dans ce corpus.

La majorité des textes soumis aux trois outils d'identification de langue ont été marqués comme non corses par ceux-ci (62,64 %), ce en quoi ils ont raison, puisque ces documents ont également été écartés par l'expert. Il existe trois configurations pour lesquelles une partie des éléments de l'échantillon correspond effectivement à

	C	M	AL	B	Total
Jugement expert	39 19,50 %	6 3,00 %	134 67,00 %	21 10,50 %	200
Identification automatique par les trois outils sélectionnés Les échantillons diffèrent par le nombre d'outils ayant reconnu le document comme corse (N). La proportion réelle de chaque combinaison dans le corpus complet est donnée entre parenthèses.					
N = 3 : Ldig + CLD3 + Réf. (5,12 % du corpus)	20 80,00 %	3 12,00 %	2 8,00 %	0 0,00 %	25
N = 2 : Ldig + CLD3 (1,98 % du corpus)	11 44,00 %	1 4,00 %	13 52,00 %	0 0,00 %	25
N = 2 : Ldig + Réf. (1,75 % du corpus)	0 0,00 %	0 0,00 %	10 40,00 %	15 60,00 %	25
N = 2 : CLD3 + Réf. (0,03 % du corpus)	8 32,00 %	1 4,00 %	16 64,00 %	0 0,00 %	25
N = 1 : Ldig (23,57 % du corpus)	0 0,00 %	0 0,00 %	25 100,00 %	0 0,00 %	25
N = 1 : CLD3 (4,33 % du corpus)	0 0,00 %	1 4,00 %	24 96,00 %	0 0,00 %	25
N = 1 : Réf. (0,57 % du corpus)	0 0,00 %	0 0,00 %	19 76,00 %	6 24,00 %	25
N = 0 (62,64 % du corpus)	0 0,00 %	0 0,00 %	25 100,00 %	0 0,00 %	25

Tableau 8. *Évaluation du corpus W2C*

du contenu corse. Lorsque CLD3 et le système par mots-clés (Réf.) identifient un contenu corse sans que Ldig ne le fasse, ce qui ne concerne que 0,03 % du corpus, le taux de précision estimé par rapport à l'échantillon est de 32 % (64 % sont dans une autre langue). La précision de l'identification augmente jusqu'à 44 % lorsque ce sont Ldig et CLD3 qui s'accordent sur la détection du corse, alors que la méthode par mots-clés donne une autre langue. Ce cas de figure, qui ne concerne que 1,98 % du corpus, laisse tout de même 52 % de textes dans une autre langue. Finalement, seule la combinaison des trois outils permet d'atteindre une identification plus fiable, avec 80 % des documents réellement en corse dans l'échantillon, 12 % de documents présentant une ou plusieurs langues en plus du corse, et 8 % de documents dans une autre langue.

L'utilisation conjointe des trois outils pourrait donc constituer une méthode de filtrage imparfaite, mais exploitable, pour ce corpus. Cette configuration, pour laquelle on peut espérer une présence au moins partielle du corse dans 92 % des cas, ne s'est cependant présentée que pour 5,12 % des documents du corpus, soit 4 629 éléments. Le taux de rappel à l'échelle du corpus serait donc assez faible, de l'ordre de 24,16 %,

et ne permettrait de conserver qu'un ensemble d'environ 4 259 documents en corse sur les 17 629 potentiellement disponibles.

5. Conclusion

L'identification de langue est un problème parfois considéré comme résolu. De nombreux points, qui s'appliquent en partie au corse, ne sont cependant pas encore maîtrisés. Il existe néanmoins des outils capables de traiter cette langue peu dotée, mais les performances ne sont pas au niveau de celles observées pour les « grandes » langues. Les résultats enregistrés sur les documents courts restent également en retrait. L'entraînement spécifique d'outils, pour le corse et une série de 17 langues européennes, a montré des performances intéressantes. Les données et procédures de génération de modèles utilisées pour cet article ont été mises à disposition[38].L'ajout de variantes dialectales, de langues proches du corse, voire d'autres langues régionales européennes, pourrait constituer une évolution importante de cet outil.

En ce qui concerne les corpus de grande envergure issus d'Internet et revendiquant du contenu corse, nous constatons une fiabilité peu élevée. Dans la lignée de Caswell *et al.* (2021), nous recommandons d'utiliser ces ressources avec prudence et de s'assurer de leur contenu par des sondages et des évaluations manuelles. Face à une ressource fortement bruitée, un filtrage automatisé ou semi-automatisé pourra, dans une certaine mesure, être mis en place au moyen de systèmes d'identification de langue. L'obtention d'une précision satisfaisante nécessite l'utilisation combinée de plusieurs d'entre eux, au prix d'un faible rappel et d'une diminution importante du volume du corpus, ce qui peut être acceptable si le corpus de départ est très volumineux. L'entraînement d'outils de filtrage spécifiques pourrait également s'avérer judicieux.

D'autre part, nous avons mentionné que la majorité de ces ressources ne prennent pas, ou peu, en compte les aspects liés aux droits d'exploitation et aux droits d'auteur.

Par conséquent, la constitution de corpus pour les langues peu dotées est loin d'être une problématique réglée, à plus forte raison si la prise en compte de la dimension dialectale est souhaitée. La création, manuelle ou automatique, de ressources textuelles fiables et sécurisées au niveau juridique reste, à notre avis, une priorité.

Remerciements

Ce travail a été mené grâce au financement CPER : « Un outil linguistique au service de la Corse et des Corses : la Banque de Données Langue Corse (BDLC) ». Nous remercions Stella Retali-Medori d'avoir consacré le temps nécessaire à l'important travail de validation des données corses. Enfin, tous nos remerciements vont aux relecteurs pour leurs commentaires avisés et constructifs.

38. Voir `https://github.com/lkevers/ldig-models-TAL62-3`, ainsi que `https://bdlc.univ-corse.fr/tal/index.php?page=lgid` pour une démonstration en ligne.

6. Bibliographie

Baroni M., Bernardini S., « BootCaT : Bootstrapping Corpora and Terms from the Web », *Proceedings of the fourth international conference on Language Resources and Evaluation (LREC 2004)*, ELRA, 2004.

Berment V., Méthodes pour informatiser les langues et les groupes de langues « peu dotées », Thèse de doctorat, Université Joseph Fourier (Grenoble), May, 2004.

Bernhard D., Bras M., Erhart P., Ligozat A.-L., Vergez-Couret M., « Language Technologies for Regional Languages of France : The RESTAURE Project », *International Conference Language Technologies for All (LT4All) : Enabling Linguistic Diversity and Multilingualism Worldwide*, Collection of Research Papers of the 1st International Conference on Language Technologies for All, ELRA, Paris, France, p. 272-275, December, 2019.

Brown R. D., « Selecting and Weighting N-Grams to Identify 1100 Languages », *Text, Speech, and Dialogue. TSD 2013*, vol. 8082, p. 475-483, 2013.

Caswell I., Breiner T., van Esch D., Bapna A., « Language ID in the Wild : Unexpected Challenges on the Path to a Thousand-Language Web Text Corpus », *Proceedings of the 28th International Conference on Computational Linguistics*, International Committee on Computational Linguistics, Barcelona, Spain (Online), p. 6588-6608, December, 2020.

Caswell I., Kreutzer J., Wang L., Wahab A., van Esch D., Ulzii-Orshikh N., Tapo A., Subramani N., Sokolov A., Sikasote C., Setyawan M., Sarin S., Samb S., Sagot B., Rivera C., Rios A., Papadimitriou I., Osei S., Suárez P. J. O., Orife I., Ogueji K., Niyongabo R. A., Nguyen T. Q., Müller M., Müller A., Muhammad S. H., Muhammad N., Mnyakeni A., Mirzakhalov J., Matangira T., Leong C., Lawson N., Kudugunta S., Jernite Y., Jenny M., Firat O., Dossou B. F. P., Dlamini S., de Silva N., Balli S. C., Biderman S., Battisti A., Baruwa A., Bapna A., Baljekar P., Azime I. A., Awokoya A., Ataman D., Ahia O., Ahia O., Agrawal S., Adeyemi M., « Quality at a Glance : An Audit of Web-Crawled Multilingual Datasets », *arXiv :2103.12028 [cs]*, April, 2021. arXiv : 2103.12028.

Cavnar W. B., Trenkle J. M., « N-Gram-Based Text Categorization », *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, p. 161-175, 1994.

Ceberio Berger K., Gurrutxaga Hernaiz A., Baroni P., Hicks D., Kruse E., Quochi V., Russo I., Salonen T., Sarhimaa A., Soria C., *Digital Language Survival Kit. The DLDP Recommendations to Improve Digital Vitality*, 2018.

Dalbera-Stefanaggi M.-J., *La langue corse*, nº 3641 in *Que sais-je ?*, PUF, Paris, June, 2002.

Dalbera-Stefanaggi M.-J., *Nouvel atlas linguistique et ethnographique de la Corse : Volume 1, Aréologie phonétique, édition revue et corrigée*, alain piazzola edn, Comité des travaux historiques et scientifiques - CTHS, Ajaccio : Paris, December, 2007.

Dunning T., Statistical Identification of Language, Technical report, 1994.

Esplà-Gomis M., Forcada M. L., Ramirez-Sanchez G., Hoang H., « ParaCrawl : Web-scale parallel corpora for the languages of the EU », p. 118-119, August, 2019.

Giguet E., « Multilingual Sentence Categorization according to Language », *Proceedings of the European Chapter of the Association for Computational Linguistics SIGDAT Workshop "From text to tags : Issues in Multilingual Language Analysis"*, p. 73-76, 1995.

Habernal I., Zayed O., Gurevych I., « C4Corpus : Multilingual Web-size Corpus with Free License », *in* N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (eds), *Proceedings of the Tenth*

International Conference on Language Resources and Evaluation (LREC 2016), ELRA, Portorož, Slovenia, May, 2016.

Hajlaoui N., Kolovratnik D., Väyrynen J., Steinberger R., Varga D., « DCEP -Digital Corpus of the European Parliament », *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, ELRA, Reykjavik, Iceland, May, 2014.

Hughes B., Baldwin T., Bird S., Nicholson J., Mackinlay A., « Reconsidering language identification for written language resources », *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006)*, ELRA, p. 485-488, 2006.

Ingle N. C., « A language identification table », *The Incorporated Linguist*, 1976.

Jaech A., Mulcaire G., Ostendorf M., Smith N. A., « A Neural Model for Language Identification in Code-Switched Tweets », *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, ACL, Austin, Texas, p. 60-64, November, 2016.

Jauhiainen T., Lui M., Zampieri M., Baldwin T., Lindén K., « Automatic Language Identification in Texts : A Survey », *arXiv :1804.08186 [cs]*, April, 2018. arXiv : 1804.08186.

Joshi P., Santy S., Budhiraja A., Bali K., Choudhury M., « The State and Fate of Linguistic Diversity and Inclusion in the NLP World », *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL, Online, p. 6282-6293, July, 2020.

Joulin A., Grave E., Bojanowski P., Mikolov T., « Bag of Tricks for Efficient Text Classification », p. 427-431, April, 2017.

Kerwin T., Classification of natural language based on character frequency, Technical report, Ohio State University, June, 2006.

Kevers L., Retali-Medori S., « Towards a Corsican Basic Language Resource Kit », *Proceedings of the 12th Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, p. 2726-2735, May, 2020.

Kevers L., Retali Medori S., Tognotti A. G., A Survey of Language Technologies Resources and Tools for Corsican, Research Report, UMR CNRS 6240 LISA, Université de Corse, 2021.

Kilgarriff A., Baisa V., Bušta J., Jakubíček M., Kovář V., Michelfeit J., Rychlý P., Suchomel V., « The Sketch Engine : ten years on », *Lexicography*, vol. 1, n° 1, p. 7-36, July, 2014.

Kornai A., « Digital Language Death », *PLoS ONE*, October, 2013.

Landragin F., *Comment parler à un alien ? : Langage et linguistique dans la science-fiction*, BELIAL, October, 2018.

Leixa J., Mapelli V., Choukri K., *Inventaire des ressources linguistiques des langues de France*, ELDA, September, 2014. Accessible à l'adresse `http://www.elda.org/media/filer_public/2014/12/17/rapport_dglflf_05112014-1.pdf`.

Lui M., Baldwin T., « Langid.Py : An Off-the-shelf Language Identification Tool », *Proceedings of the ACL 2012 System Demonstrations*, ACL, Stroudsburg, PA, USA, p. 25-30, 2012. event-place : Jeju Island, Korea.

Majliš M., « Yet Another Language Identifier », *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, ACL, Avignon, France, p. 46-54, April, 2012.

Majliš M., Zabokrtský Z., « Language Richness of the Web », *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, Istanbul, Turkey, May, 2012.

Marcellesi J.-B., « La définition des langues en domaine roman : les enseignements à tirer de la situation corse », *Actes du Congrès de Linguistique et de Philologie Romanes 5*, Aix-en-Provence, p. 307-314, 1984.

Millour A., Myriadisation de ressources linguistiques pour le traitement automatique de langues non standardisées, Thèse de doctorat, Sorbonne Universite, December, 2020.

Moseley C. (ed.), *Atlas of the World's Languages in Danger*, UNESCO Publishing, Paris, 2010. 3rd edn. `http://www.unesco.org/culture/en/endangeredlanguages/atlas`.

Nakatani S., « Language Detection Library for Java », March, 2010. `https://www.slideshare.net/shuyo/language-detection-library-for-java`.

Nakatani S., « Short Text Language Detection with Infinity-Gram », May, 2012. `https://www.slideshare.net/shuyo/short-text-language-detection-with-infinitygram-1294944`.

Okanohara D., Tsujii J., « Text Categorization with All Substring Features », *Proceedings of SDM 2009*, p. 838-846, April, 2009.

Rehurek R., Kolkus M., « Language Identification on the Web : Extending the Dictionary Method », *n : CICLing '09 : Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing*, p. 357-368, March, 2009.

Roziewski S., Stokowiec W., « LanguageCrawl : A Generic Tool for Building Language Models Upon Common-Crawl », *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, ELRA, Portorož, Slovenia, p. 2789-2793, May, 2016.

Scannell K. P., « The Crúbadán Project : Corpus building for under-resourced languages », *in* C. Fairon, H. Naets, A. Kilgarriff, G.-M. de Schryver (eds), *Proceedings of the 3rd Web as Corpus Workshop*, vol. 4 of *Cahiers du Cental*, Louvain-la-Neuve, Belgium, 2007.

Siewert J., Scherrer Y., Wieling M., Tiedemann J., « LSDC - A comprehensive dataset for Low Saxon Dialect Classification », *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, International Committee on Computational Linguistics (ICCL), Barcelona, Spain (Online), p. 25-35, December, 2020.

Soria C., Mariani J., Zoli C., « Dwarfs sitting on the giants' shoulders – how LTs for regional and minority languages can benefit from piggybacking major languages », *XVII FEL Conference*, Ottawa, October, 2013.

Suárez P. J. O., Romary L., Sagot B., « A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages », *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 1703-1714, July, 2020.

Tahir B., Mehmood M. A., « Corpulyzer : A Novel Framework for Building Low Resource Language Corpora », *IEEE Access*, vol. 9, p. 8546-8563, 2021.

Takçı H., Güngör T., « A high performance centroid-based classification approach for language identification », *Pattern Recognition Letters*, vol. 33, n° 16, p. 2077-2084, December, 2012.

Wenzek G., Lachaux M.-A., Conneau A., Chaudhary V., Guzmán F., Joulin A., Grave E., « CCNet : Extracting High Quality Monolingual Datasets from Web Crawl Data », *arXiv :1911.00359 [cs, stat]*, November, 2019. arXiv : 1911.00359.

Xue L., Constant N., Roberts A., Kale M., Al-Rfou R., Siddhant A., Barua A., Raffel C., « mT5 : A Massively Multilingual Pre-trained Text-to-Text Transformer », *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, ACL, Online, p. 483-498, June, 2021.

Towards a Low-Resource Neural Machine Translation for Indigenous Languages in Canada

Ngoc Tan Le* — **Fatiha Sadat***

* *Department of Computer Science, University of Quebec in Montreal, Canada*

ABSTRACT. *The Natural Language Processing research community is increasingly interested in less-resourced languages and linguistic diversity through technology. Translation to and from low-resource polysynthetic languages has, in particular, always faced numerous challenges, such as morphological complexity, dialectal variations, noisy data due to different spellings and low-resource scenarios. Moreover, the morphological segmentation for indigenous polysynthetic languages is particularly challenging with multiple individual morphemes by word and several meanings per morpheme. The present research focuses on Inuktitut and Inuinnaqtun, indigenous polysynthetic languages spoken in Northern Canada. We then build a morphological segmenter and a NMT system for these indigenous languages. Our proposed NMT model outperformed the state-of-the-art in the context of low-resource Inuktitut-English Neural Machine Translation.*

RÉSUMÉ. *La communauté de recherche sur le traitement des langues naturelles porte un intérêt croissant aux langues peu dotées et à la diversité linguistique grâce à la technologie. La traduction vers et depuis les langues polysynthétiques s'est régulièrement heurtée à de nombreux défis comme la complexité morphologique, les variants dialectiques, les données bruitées, les différentes orthographes, et les scénarios d'entraînement avec peu de données. Par ailleurs, la segmentation morphologique des langues polysynthétiques autochtones est rendue particulièrement difficile en raison de multiple morphèmes par mot et de plusieurs sens par morphème. La présente recherche se concentre sur l'inuktitut et l'inuinnaqtun, langues polysynthétiques autochtones parlées dans le nord du Canada. Nous construisons un segmenteur et un système de traduction automatique neuronale pour langues autochtones du Canada. Notre modèle de traduction automatique a surpassé l'état de l'art dans le contexte de la traduction automatique neuronale inuktitut-anglais.*

KEYWORDS: *Polysynthetic languages, Inuktitut, Inuinnaqtun, NMT, Low-resource.*

MOTS-CLÉS : *Langues polysynthétiques, Inuktitut, Inuinnaqtun, TAN, peu dotée.*

1. Introduction

According to Mager *et al.* (2018), the Americas have a diverse range of linguistic families, with approximately 900 different indigenous languages spoken. More specifically, Canada's wide range of indigenous languages, grouped into 12 language families, has played an important role in the history of First Nations, Métis, and Inuit, and continues to do so today (Rice, 2011). Due to a variety of factors, there has been very little research on indigenous languages in recent years. Natural Language Processing (NLP) researchers working with indigenous languages encounter numerous obstacles, including polysynthesis, with a high rate of morphemes per word, lack of orthographic normalization, dialectal variances, and a lack of linguistic resources and tools (Littell *et al.*, 2018; Schwartz *et al.*, 2020).

This study focuses on two indigenous polysynthetic languages spoken in Northern Canada, particularly Inuktitut and Inuinnaqtun, as well as the development of an Inuktitut-English Neural Machine Translation (NMT).

In the Northwest Territories, Inuktitut and Inuinnaqtun (a related dialect group) are recognized as official indigenous languages. They belong to the language family of Esquimo-Aleut, including the Inuit language. The Inuit language, or Inuktut, is a continuum of dialects that are spoken in the North American Arctic: in northern Alaska, in the Northwest Territories, in Nunavut, in Nunavik (northern Quebec), Nunatsiavut (in Labrador), and Greenland. Inuktitut [1] is an indigenous North American language spoken in the Canadian Arctic. Inuktitut is part of the vast Inuit language continuum (set of dialects) stretching from Alaska to Greenland. Inuktitut has official language status in Nunavut, like English and French. According to the 2016 census [2], it has approximately 39,770 speakers, 65% of whom live in Nunavut and 30.8% in Quebec. Inuinnaqtun belongs to the Western Canadian Inuktun family of languages, including two other dialects, Siglitun and Natsilingmiutut. According to Statistics Canada, in 2016, Inuinnaqtun is the mother tongue of 675 people in Canada and 1,310 people can speak this language.

This first step towards a multilingual NMT framework, which will include several endangered indigenous languages of Canada, is critical, the Nunavut-English Hansard corpus being the only parallel corpus freely available for research (Joanis *et al.*, 2020). Haddow *et al.* (2021) considered many features between high- (*e.g.* 280M parallel sentences), medium- (*e.g.* 0.7M parallel sentences), and low-resource (*e.g.* 0.035M parallel sentences) language pairs based on the number of native speakers and the quantity of parallel sentences. Joshi *et al.* (2020) presented the relationships between the types of languages, resources, and their representation in NLP conferences to understand the trajectory that different languages have followed over time. They highlighted, via a quantitative investigation, the disparity between languages, especially

1. Source: Compton, Richard . "Inuktitut". L'Encyclopédie Canadienne, 20 novembre 2019, Historica Canada. `www.thecanadianencyclopedia.ca/fr/article/inuktitut`.

2. Source: `https://www12.statcan.gc.ca/census-recensement/2016/as-sa/98-200-x/2016022/98-200-x2016022-fra.cfm`.

in terms of their resources. Although there is a correlation between them, there are many outliers where either widely spoken languages have a minimal parallel corpus, or languages with a limited number of speakers are resource-rich in terms of corpora. We also observed that the concept of low-resource might change over time. We could crawl additional parallel data, or use related language data or monolingual data. Several language pairings are no longer considered low-resource. Thanks to the crawled parallel sentences size, Inuktitut is currently rated a medium-resource level. However, with only a few comparable phrases, such as the Bible, Inuinnaqtun remains highly under-resourced.

The primary goal and motivation for this research project aim to revitalize and to preserve Canadian indigenous languages and cultural heritage through major NLP tasks. Our research is divided into two stages: (1) building a morphological segmenter for indigenous languages, to be integrated into (2) the framework of a Neural Machine Translation system for indigenous languages.

Inspired by the work of Farley (2012), related to the creation of the first Inuktitut finite-state transducer-based morphological analyzer, we propose a novel technique based on deep neural networks to create a word segmenter for indigenous languages. First, we investigate several methods empirically, including supervised, semi-supervised and non-supervised approaches to word segmentation task. In the supervised approach, the task is considered as a sequence labelling task. We apply the sequence-to-sequence architecture (Sutskever *et al.*, 2014) with the encoder-decoder layers (see Section 3.1). In the semi-supervised and non-supervised approaches, we adopt an Adaptor Grammars (Johnson, 2008), fine-tuning the word segmentation model using a deep learning-based architecture for indigenous languages (see Section 3.2). Second, we construct a framework for a low-resource Neural Machine Translation system by incorporating our word segmenter, during the source-side language preprocessing step (see Section 3.5).

Our contributions to the current research are as follows:

(1) to perform empirical research on several word segmentation approaches to indigenous languages, particularly Inuktitut and Inuinnaqtun;

(2) to propose a neural network-based word segmenter for indigenous languages;

(3) to enhance low-resource NMT via extensive morphological word segmentation;

(4) to empirically compare our proposed NMT technique with different designs such as Sequence-to-Sequence (Sutskever *et al.*, 2014), Transformer (Vaswani *et al.*, 2017), and multilingual NMT architecture.

The following is a description of the article's structure: The section 2 highlights the most recent advances in morphological analyzers and Machine Translation concerning indigenous languages. Our technique is described in Section 3. Section 4 provides our experiments and results. Finally, Section 5 offers our conclusion as well as potential future research directions.

2. Related work

2.1. *Morphological complexity of indigenous languages*

Most of the indigenous languages, particularly in the Americas, belong to either the polysynthetic language group or the agglutinative language group. They have a complex, rich morphology that plays an important role in human learning versus machine learning (Gasser, 2011; Littell *et al.*, 2018). Much of the research on their morphological analysis has focused only on linguistic aspects.

From a linguistic point of view, a word is formed from a word base, followed by multiple suffixing morphemes, to become a long sentence word. The suffixing morphology phenomenon makes a word very long and potentially unique (Mithun, 2015). Moreover, the composition of verbs and the incorporation of noun in indigenous languages are much more complex than in European languages, such as English or French. Indigenous languages of North America are highly inflected.

Our current research investigates Inuktitut and Inuinnaqtun, polysynthetic languages in the Inuit language family spoken in Northern Canada. Inuinnaqtun is endangered and has been estimated to disappear in less than two generations[3].

Comparing word composition in English, the word structure in Inuit languages is variable in its surface form. Words can be very short, composed of three formative features such as word base, lexical suffixes, and grammatical ending suffixes. They can be very long up to ten or even fifteen formative morphemes as features depending on the regional dialect (Lowe, 1985; Kudlak and Compton, 2018).

Composition of Inuit word = **Word base** + Lexical suffixes + *Grammatical ending suffixes*

For instance, in the Inuktitut language of Nunavut[4], Canada, a sentence word example is given below:

– (Inuktitut script) ᑐᓵᑦᓯᐊᕈᓐᓇᖏᑦᑐᐊᓘᔪᖓ

– (Romanized script) **tusaa**-tsia-runna-nngit-tu-alu-u-*junga*

– (Meaning) **hear**-well-be.able-NEG-DOER-very-*BE-PART.1.S*, where NEG-DOER means the negation, and PART.1.S means participle first singular.

– (English) I can't hear very well

In this example, this long word is composed of the morpheme as a root word **tusaa** (meaning: **to hear**), followed by seven morphemes: six lexical suffixes (**tsia, runna, nngit, tu, alu, u**) and one grammatical ending suffix (**junga**) that are separated by hyphens. All segmented morphemes are synthetically composed in only one unit.

3. Source: `https://www.kitikmeotheritage.ca/language`.
4. Source: `https://en.wikipedia.org/wiki/Inuit_grammar`.

In Inuinnaqtun, another sentence word example is depicted from the grammar book of Lowe (1985), illustrating the same phenomenon of word composition of an Inuinnaqtun sentence word *umingmakhiuriaqtuqatigitqilimaiqtara* that can be segmented into a word base and several suffixes as follows:

– (Inuinnaqtun script) umingmakhiuriaqtuqatigitqilimaiqtara

– (Morpheme segmentation) **umingmak**-hiu-riaqtu-qati-gi-tqi-limaiq-*ta-ra*

– (Meaning) **muskox** - hunt - go in order to - partner - have as - again - will no more - *I-him*

– (English) I will no more again have him as a partner to go hunting **muskox**

In this example, the first morpheme as a root word **umingmak** (meaning: **muskox**) is followed by six morphemes as lexical suffixes (**hiu, riaqtu, qati, gi, tqi, limaiq**) and two grammatical ending suffixes (**ta, ra**).

A single word can be used to express what would be a whole sentence in English. We note that the word composition tends generally to augment the lexical constituents with multiple formative suffixing morphemes added to a word base. Full sentences are commonly made up of only one word. Moreover, the morphology is highly inflected with a variety of lexical suffixes and grammatical ending suffixes. All these linguistic aspects make the morphological segmentation task for polysynthetic languages more challenging. One of the challenges consists in determining the word that is the basic unit, then the sub-word units (Arppe *et al.*, 2017).

2.2. *Morphological segmentation of indigenous languages*

The development of a morphological segmenter for indigenous languages was not well supported due to several challenges, as indicated above. Unsurprisingly, owing to the lack of annotated data, we used an unsupervised approach, as well as the rule-based approach used numerous works. Creutz and Lagus (2007) proposed the statistical morphological segmentation method, named Morfessor, based on the Hidden Markov Model for learning unsupervised morphology, and using a hierarchical morpheme structure.

Another method shown to be successful for unsupervised morphological segmentation is the Adaptor Grammars (AG) approach, based on non-parametric Bayesian models generalizing probabilistic context-free grammar (PCFG) (Johnson, 2008). By defining a set of morphological grammar patterns, including zero or more prefixes, stems, and suffixes, the AG models are able to induce segmentation at the morpheme level. Several studies have been conducted based on extending this approach, such as those of Botha and Blunsom (2013) for learning non-concatenative morphology, Sirts and Goldwater (2013) for minimally supervised morphological segmentation, and Eskander *et al.* (2018) for unseen languages. Godard *et al.* (2018) used this approach to experiment with the word segmentation task in very low-resource African languages. Eskander *et al.* (2019) also used this approach to deal with Mexican low-resource

polysynthetic languages such as Mexicanero, Nahuatl, Yorem Nokki and Wixarika. In the current work, we also examine the efficiency of the AG-based approach on the Inuktitut language, a polysynthetic low-resource language without annotated segmented resources.

In terms of the Inuktitut language, we noted only a few studies on morphological segmentation task. Johnson and Martin (2003) proposed an unsupervised technique, with the hubs concept in a finite-state automaton. The hubs mark the boundary between root and suffix. Concretely, Inuktitut words are segmented into morphemes and merged hubs in a finite-state automaton. They reported good performance for English morphological analysis, using the text of *Tom Sawyer*, for which they obtained 92.15% in terms of precision. However, for Inuktitut morphological analysis, they reported 31.80% precision and a low recall of 8.10%. They argued the poor performance for Inuktitut roots was due to the difficulty of identifying word-internal hubs. Farley (2012) proposed hand-crafted grammar rules and a finite-state transducer to build a morphological analysis for Inuktitut, called *Uqailaut* (pronounced Uqa·Ila·Ut). This Uqailaut project is a rule-based system based on regular morphological variations of about 3,200 head (or prefix), 350 lexical, and 1,500 grammatical morphemes, with heuristics for ranking the various readings. Nicholson *et al.* (2012) used a word alignment error rate with the dataset of English-Inuktitut Nunavut parallel corpora to evaluate the morphological analyzer for Inuktitut. They reported their best experimental results, in terms of the head (or prefix) approach, which, in Inuktitut, corresponds to the first one or two syllables of a token, with 79.70% precision and 92.20% recall. They reported that the analyzer was able to provide at least a single analysis for approximately 218k Inuktitut types (65%) from the Nunavut Hansard corpus. In addition, Micher (2017), inspired by Farley (2012) Uqailaut project, used a segmental recurrent neural network approach based on the output of this morphological analyzer for Inuktitut. The models were trained with approximately 23k types having a single analysis from the Uqailaut analyzer, with 85.07% in terms of F-measure.

2.3. *Machine translation for indigenous languages*

Machine translation (MT) is well known in language technologies. Building a reliable, high quality MT system is still a significant challenge for indigenous languages. Mager *et al.* (2018) reported an interesting and ongoing research problem in the MT task of low-resource languages, especially indigenous languages. We reviewed the development of MT systems for indigenous languages based on the following fundamental approaches: (1) rule-based, (2) statistics-based, and (3) neural network-based approaches.

(1) Rule-Based Machine Translation (RBMT) approaches are usually applied in the low-resource languages scenario. The RBMT systems do not require aligned parallel corpora. However, they require language-dependent knowledge. They have several drawbacks, mostly pertaining to translating complex structures and to building

complex rules. Apertium [5] is a free and open-source platform for developing rule-based machine translation systems. Recently, research on data-driven approaches has improved to deal with data scarcity and data sparsity (Mager *et al.*, 2018).

(2) In Statistical-based Machine Translation (SMT) approaches, for translating to and from morphologically complex languages, researchers have proposed treating words as sentences or subword units. The performance of SMT systems is highly dependent on the quantity of training data, which represents a challenge when dealing with low-resource conditions. In the case of the native languages of the Americas, the SMT systems were challenged by the rich and complex morphology and the data sparseness (Micher, 2017) of the languages. We examined a variety of applications of this research and its foundation in the SMT line of research. Sennrich *et al.* (2016) proposed using byte pair encoding (BPE) to segment words into subword units and showed improvement in machine translation on an English to German and English to Russian task of up to 1.1 and 1.3 BLEU, respectively. Micher (2018) reported 30.04 BLEU in the English to Inuktitut direction, and 30.35 BLEU in the Inuktitut to English direction, using the BPE-preprocessed the Nunavut Hansard Inuktitut-English parallel corpora.

(3) Neural network-based Machine Translation (NMT) approaches use neural networks architectures trained with vast amounts of parallel texts. In this approach, the NMT systems are applied in several neural networks architectures such as Seq2Seq (Sutskever *et al.*, 2014), Transformer with Encoder-Decoder and Attention (Vaswani *et al.*, 2017). These systems work well when dealing with resource-rich language pairs because the training requires a significant quantity of parallel texts.

In Machine Translation task for indigenous languages, several NMT systems were presented at the WMT 2020 workshop [6] for multiple languages pairs, including Inuktitut-English. We compare our NMT model against some of them in the sub-Section 4.3.

Building an MT system for indigenous languages is considered a low-resourced scenario (Schwartz *et al.*, 2020). For many low-resourced language pairs, the corpora are derived from religious sources (*e.g. the Bible or Koran*) or technical documents (*e.g. Opus* (Tiedemann, 2012)), or from IT data localization (e.g. from open-source projects such as *GNOME or Kubuntu*) (Haddow *et al.*, 2021). Recently, Nicolai *et al.* (2021) built the JHU Bible corpus for MT of the indigenous languages of North America, with $26k$ verses in the Inuktitut family language, and it achieved only 11.8 in terms of BLEU score. Joanis *et al.* (2020) constructed $1.29M$ bilingual sentences in the Nunavut Hansard for Inuktitut-English (third edition), available for research purposes. Research on NMT with low-resource language pairs still face multiple compounding major challenges, such as lack of NLP tools, lack of parallel corpora, out-of-domain data, and noisy data (Littell *et al.*, 2018; Mager *et al.*, 2018; Joanis *et al.*, 2020; Le and Sadat, 2020; Mager *et al.*, 2021). Aside from data problems, indigenous languages are

5. Apertium: `https://wiki.apertium.org/wiki/Main_Page`.
6. Source: `http://www.statmt.org/wmt20/translation-task.html`.

frequently understudied languages, in which access to local speakers and specialists is difficult, and even fundamental toolkits such as language identifiers or morphological analyzers do not exist or are not trustworthy (Haddow *et al.*, 2021).

3. Proposed methodology

3.1. *Supervised approach for the morphological segmentation*

The neural network-based approach can be efficiently applied on word segmentation using pretrained embeddings and several deep learning techniques. Furthermore, using additional linguistic factors helps the neural model perform better, especially when dealing with data sparseness or language ambiguity in the context of indigenous languages (Kann *et al.*, 2018).

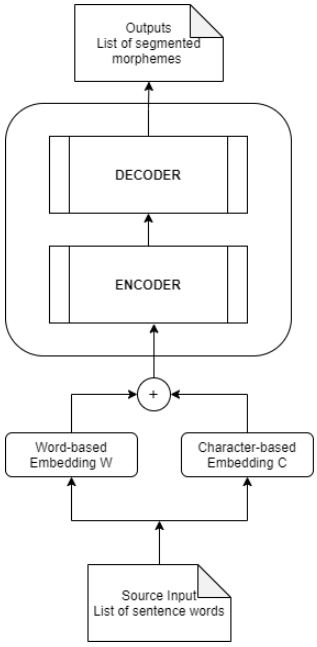

Figure 1. *Architecture of our framework: Morphological segmentation for indigenous language based on the encoder-decoder architecture.*

The goal of morphological segmentation is to divide words into morphemes. This task may be thought of as a structured classification problem, with each character being allocated to one of many predefined classes. These classes are denoted as follows: (B) represents the beginning of a multi-character morpheme, (M) the middle of a multi-character morpheme, and (E) the end of a multi-character morpheme,

and (S) denotes a single character morpheme. Other schemes are also conceivable such as IOB format (short for inside, outside, beginning) or IO or BMEO or BM (Carpenter, 2009; Ruokolainen *et al.*, 2013; Wang *et al.*, 2016).

For instance, for the Inuktitut word "*tusaattialaurit*" (meaning: to listen), the corresponding morphological segmentation should be:

tusaa+tti+ala+u+rit

By adding the two extra symbols <w> and </w> to indicate the start and the end of a word, respectively, the above segmentation form is represented as follow:

<w> tusaa tti ala u rit </w>
START BMMME BME BME S BME STOP

In this research, the morphological segmentation task is considered as a sequence labeling task, with the goal of classifying each character in a word into the appropriate class. Given an input sequence, $W = [w_0, w_1, ..., w_m]$ and $C = [c_0, c_1, ..., c_n]$ contain all the input words and the input characters. The architecture is based on Sequence-to-Sequence model (Sutskever *et al.*, 2014) with the encoder-decoder layers as shown in Figure 1. The encoder layer contains the input word sequence transformation by concatenating pretrained character-based and word-based embeddings, with the state $S = \langle W, C \rangle$. We apply the attention mechanism that allows the model to focus, in the context, on a set of characters and to learn the important letters to better predict whether a character forms a boundary. We introduce an attention vector, a_i, used to measure the weight of the sentence words in the context. The resulting context embedding, v_i, jointly learned during the training phase, helps to capture the relevant information from the context.

$$\alpha_t = \frac{exp(score(h_t, h_{s}))}{\sum_{s'=1}^{S} exp(score(h_t, h_{s}))} \quad [1]$$

$$c_t = \sum_s \alpha_t h_s \quad [2]$$

$$a_t = f(c_t, h_t) = tanh(W_c[c_t; h_t]) \quad [3]$$

where α_t is the attention weight of the target words in the context, h_s and h_t are the weight of the hidden layer for source and target words, respectively, c_t is the context vector and a_i is the attention vector.

The decoder layer $P(y|s_i)$ calculates the activation function θ as an output function and displays the output hypothesis, where y is the output prediction, s_i is a word sentence, and b_i is a bias.

$$P(y|s_i) = \theta(W_h \cdot a_i + b_i) \quad [4]$$

3.2. *Unsupervised approach for the morphological segmentation*

Inspired by the work of Eskander *et al.* (2019), we adapted an unsupervised approach in learning all possible morphological patterns using Adaptor Grammars (Johnson, 2008) and fine-tuned the outputs of the first stage by building a recurrent neural network-based architecture for Inuktitut. This approach is, basically, based on the grammar containing production rules, non-terminal and terminal symbols, and a lexicon. The deep learning can successfully handle data sparsity or language ambiguities thanks to additional linguistic factors related to a set of rich, high-quality features, such as semantic distribution information, and contextual meanings that are extracted from pretrained embeddings at character level and at word level, learned from monolingual large-scale raw corpora (Kann *et al.*, 2018).

The first phase in Adaptor Grammars-based learning consists of defining the grammar, including non-terminals, terminals, and production rules. As explained in Eskander *et al.* (2019), the grammar construction relies on three main dimensions: word modeling, abstraction level, and segmentation boundaries. The grammar patterns specify the word structures where a word is considered a sequence of prefixes, a stem, and a sequence of suffixes. Moreover, each production rule has two parameters to configure, a and b, in the Pitman-Yor process (Pitman and Yor, 1997). Setting $a = 1$ and $b = 1$ indicates to the running learner that the current non-terminals are not adapted and sampled by the general Pitman-Yor process. Otherwise, the current non-terminals are adapted and expanded as in a regular probabilistic context-free grammar. The standard grammar setting (Table 1) is language-independent, and contains all possible generic patterns, whereas the scholar-seeded grammar setting (Table 2) combines all standard grammar patterns and additional language-dependent knowledge, in this case a list of affixes. By using the list of affixes and roots, called Scholar-seeded setting, we inject linguistic knowledge into the training phase. Then, we still apply the probabilistic context-free grammar (PCFG). The model is, therefore, able to learn more patterns, with non-concatenative morphology, and to induce segmentation at the morpheme level.

We fine-tuned the outputs of the first stage through a Recurrent Neural Network-based (RNN) architecture. These outputs are fed into a bidirectional Long-Short Term Memory (Hochreiter and Schmidhuber, 1997). Formally, these input sequences are numerically vectorized using pretrained embeddings, at word-level W and at character-level C representations. The hidden feature layer then merges all input features X_W, X_C in a single vector with a k-dimension, $\langle W, C \rangle$. The output layer calculates an activation function θ, where W_o is the output weight, h is the hidden layer, and b_o is its bias.

$$h = tanh(W_{hW} \cdot X_W + W_{hC} \cdot X_C) \quad [5]$$

$$output = \theta(W_o \cdot h + b_o) \quad [6]$$

	Suffix –> $$$
1 1 Word –>Prefix Stem Suffix	Suffix –>SuffixMorphs $$$
	1 1 SuffixMorphs –>SuffixMorph SuffixMorphs
Prefix –>^^^	1 1 SuffixMorphs –>SuffixMorph SuffixMorph –>SubMorphs
Prefix –>^^^PrefixMorphs	1 1 SubMorphs –>SubMorph SubMorphs
1 1 PrefixMorphs –>PrefixMorph PrefixMorphs	1 1 SubMorphs –>SubMorph
1 1 PrefixMorphs –>PrefixMorph	SubMorph –>Chars
PrefixMorph –>SubMorphs	1 1 Chars –>Char
	1 1 Chars –>Char Chars
Stem –>SubMorphs	

Table 1. *The standard PrefixStemSuffix+SuffixMorph grammar for Inuktitut. The symbols ^^^and $$$ mean the beginning and the end of the word sequence, respectively. Source: Eskander* et al. *(2019).*

[All standard setting grammar in Table 1]
1 1 PrefixMorph –> (a) (u) (l) (l) (a)
1 1 PrefixMorph –> (i) (g) (l) (u)
1 1 PrefixMorph –> (q) (i) (n) (m) (i)
1 1 PrefixMorph –> (u) (t) (a) (q) (q) (i)
[...]
1 1 SuffixMorph –> (a) (n) (n) (i) (n)
1 1 SuffixMorph –> (f) (f) (a) (a) (n) (g) (m) (i)
1 1 SuffixMorph –> (g) (i) (a) (q) (t) (u) (q)
1 1 SuffixMorph –> (m) (i) (u) (t) (a) (t)
1 1 SuffixMorph –> (n) (') (n) (g) (u) (l) (i) (q)
1 1 SuffixMorph –> (y) (u) (t)
[...]
1 1 Char –> (q)
1 1 Char –> (k)
[...]
1 1 Char –> (p)
1 1 Char –> (t)

Table 2. *The scholar-seeded PrefixStemSuffix+SuffixMorph grammar for Inuktitut, with prefixes, suffixes, and characters.*

3.3. *Uqailaut morphological analyzer for Inuktitut*

The Uqailaut project, proposed by Farley (2012), is based on a Finite-State Transducers (FST), while applying several techniques and resources such as grammar rules, linguistic knowledge and heuristics. The FST-based morphological analyzer produces one or more morphological predictions for a given word. Heuristics make it possible to choose the shortest path for the morphological analysis. For example, *tusaattialau-*

rit is segmented as ***tusaa*** *tti ala u rit* or ***tusaa*** *ttia lau rit* or ***tusaa*** *ttia la u rit* (Table 3). The root ***tusaa*** means *to listen*, *tti*, *ala*, *u* are lexical suffixes, and *rit* is a grammatical suffix.

Morphological Segmentation	**Output**
Raw text	tusaattialaurit
Reference	**tusaa** tti ala u *rit*
First best prediction	**tusaa** tti ala u *rit*
Second best prediction	**tusaa** ttia lau *rit*
Third best prediction	**tusaa** ttia la u *rit*

Table 3. *Predictions of the Inuktitut morphological segmentation by the Uqailaut analyzer (Meaning: please listen).*

3.4. *Byte-Pair Encoding segmentation*

Sennrich *et al.* (2016) proposed the Byte-Pair Encoding (BPE) method for the word segmentation task. This method consists of unsupervised word segmentation that tries to break words into subword units, which aids in dealing with unusual and unfamiliar terms.

BPE uses the minimum entropy on subword units, often known as tokens, with a given vocabulary size. Although these tokens resemble morphemes, the BPE segmentation model is based on training data rather than linguistic knowledge bases.

For example, in Inuktitut, *'tusaattialaurit*' (meaning: *please listen* in English) may be segmented as *'tusaa@@ tti@@ alau@@ rit'* (Table 4). This word should be correctly segmented *'tusaa@@ tti@@ ala@@ u@@ rit'*, in the case of large-scale training data.

Method	**Morphological Segmentation Output**
Raw text	tusaattialaurit
Reference	tusaa tti ala u rit
Uqailaut analyzer (Farley, 2012)	tusaa tti ala u rit
BPE (Sennrich *et al.*, 2016)	tusaa@@ tti@@ ala@@ u@@ rit
Our proposed approach	tusaa tti ala u rit

Table 4. *Illustration of several Inuktitut word segmentation methods. The symbol @@ represents an in-word morpheme boundary (Meaning: please listen).*

3.5. *Polysynthetic indigenous language NMT*

The second phase of our framework consists of building an NMT for indigenous language to English based on the Transformer encoder-decoder architecture (Vaswani *et al.*, 2017). We apply our morphological segmentation method to preprocess the source indigenous language in the context of an Inuktitut-English NMT system.

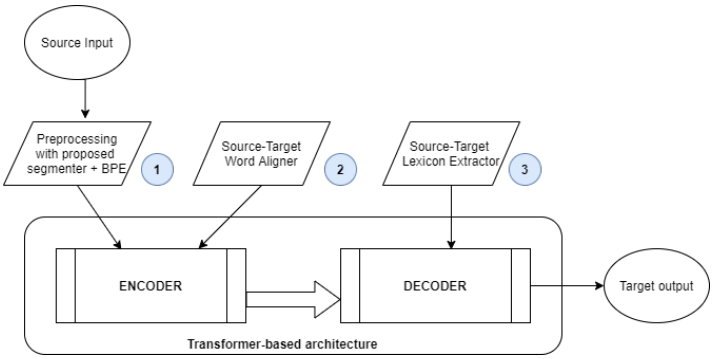

Figure 2. *Architecture of our Inuktitut-English NMT with three main parts: (1) Preprocessing with our proposed morphological segmenter and Byte Pair Encoding (BPE) for both source and target languages, respectively, (2) Building a source-target word aligner and (3) Building a source-target lexicon extractor.*

3.5.1. *Word Aligner and Lexicon Extractor for NMT*

This architecture aims to investigate our Inuktitut-English neural machine translation system while using the word alignment information and the source-target lexicon. Our approach consists of three main parts. First, the source input is preprocessed by applying our proposed morphological segmenter and Byte Pair Encoding (BPE) (Sennrich *et al.*, 2016) for both source and target languages, respectively. Second, the word alignment information is extracted from the bilingual parallel corpus and is fed into the encoder. Third, we prepare a bilingual source-target lexical shortlist. This bilingual lexicon is then used during the decoding.

3.5.2. *Multilingual NMT architecture*

Adding data from multiple languages, *i.e.* multilingual NMT, can enhance the performance of NMT systems (Aharoni *et al.*, 2019). We adapt this approach in the context of low-resource indigenous languages by using several closely-related languages (Figure 3).

For each language pair, a BPE-based model is learned jointly from the source-target sides of the parallel corpora using *subword-nmt* (Sennrich *et al.*, 2016). In addition, the source-side indigenous languages, here Inuktitut and Inuinnaqtun, are segmented by our proposed word segmenter. Then the joint BPE model is applied to all the training datasets. Moreover, we apply the BPE drop-out (Provilkov *et al.*, 2019) to deal with data sparsity and morphological complexity, such as orthographic variation or spelling errors.

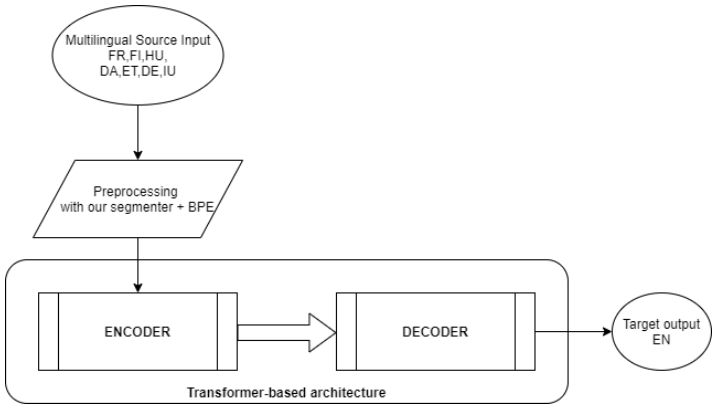

Figure 3. *Architecture of our multilingual NMT system. Here, multilingual source input is composed of multiple related languages such as French-fr, Finnish-fi, Danish-da, German-de, Hungarian-hu, Estonian-et and Inuktitut-iu, and the target output is English-en.*

4. Experiments

4.1. *Data preparation*

In our experiments and evaluations, we used the third edition of the Inuktitut-English Nunavut Hansard (Joanis *et al.*, 2020) to train our models. This parallel corpus contains 1,293,348 training sentences, 5,433 development sentences and 6,139 testing sentences, respectively. Furthermore, in order to develop our multilingual NMT model, we used several parallel corpora, including multiple language sources with an English target, provided from the shared task of WMT 2020. Tables 5 and 6 describe the statistics of the training corpora.

	#tokens	**#train**	**#dev**	**#test**
IU	20,657,477	1,293,348	5,433	6,139
EN	10,962,904	1,293,348	5,433	6,139

Table 5. *Statistics of the Nunavut Hansard for Inuktitut-English.*

Inuktitut corpus is transformed from syllabic to roman using the *uniconv* toolkit [7]. We then apply consistent preprocessing with English defaults on both source and target languages of the parallel corpora using Moses (Koehn *et al.*, 2007) scripts such as punctuation normalization, tokenization, cleaning the training corpus and truecasing on the training datasets.

7. Toolkit *uniconv* with Yudit: `www.yudit.org`.

Source-Target [English]	#train	#dev	#test
Finnish (fi-en)	1,918,232	1,000	–
French (fr-en)	2,002,165	1,000	–
Hungarian (hu-en)	623,448	1,000	–
Danish (da-en)	1,949,393	1,000	–
Estonian (et-en)	651,746	1,000	–
German (de-en)	1,920,209	1,000	–
Inuktitut (iu-en)	1,293,348	5,433	6,139
Inuinnaqtun (ikt-en)	3,511	–	–

Table 6. *Statistics of all corpora for multilingual NMT model training (Source: WMT 2020). Inuinnaqtun corpus is extracted from the Nunavut Government Website: https://www.gov.nu.ca/in/.*

For the Inuinnaqtun dataset, we manually collected a small corpus from several resources such as the Nunavut Website [8] government, open source dictionaries and grammar books (Lowe, 1985; Kudlak and Compton, 2018). The experimental corpus contains 190 word bases and 571 affixes. A small golden testing set was manually crafted containing 1,055 unique segmented words.

4.2. *Training settings*

To train the supervised morphological segmentation model, we adapted the *RichWordSegmenter* toolkit (Yang *et al.*, 2017). We chose Inuktitut source from the Nunavut Hansard to perform experiments. Then, using the Uqailaut toolkit (Farley, 2012), we annotated $11k$ training sentences, 250 development sentences, and 250 testing sentences. To pre-train the character-based and word-based embeddings, we used the Nunavut Hansard Inuktitut corpus 3.0 and the *Gensim* [9] library to train all embeddings with a dimension of 30 and 50, respectively. We found that there are only 97,785 unique terms for the word-based vocabulary, 102 unique terms for the character-based vocabulary and 1,406 unique terms for the character-based vocabulary (Table 7).

Embedding type	#terms	#dimension
word-based	97,785	50
character-based	102	30
bicharacter-based	1,406	30

Table 7. *Statistics of word-based and (bi)character-based embeddings training using Nunavut Hansard Inuktitut-English parallel corpus 3.0 for Inuktitut.*

8. Nunavut government Website in Inuinnaqtun: `https://www.gov.nu.ca/in/cgs-in`.
9. Gensim library: `https://radimrehurek.com/gensim/models/word2vec.html`.

The two principal inputs, used to train the Adaptor Grammars-based unsupervised morphological segmentation model, consist of the grammar and the lexicon of the language. The learning hyperparameters are configured as in Eskander *et al.* (2019) according to the best standard PrefixStemSuffix+SuffixMorph grammar (Table 1) and the best scholar-seeded grammar (Table 2). Next, we fine-tuned the outputs of the first stage with an RNN-based architecture consisting of bi-directional Long Short-term Memory, with 200 neurons in the hidden layer. We evaluated both supervised and unsupervised proposed morphological segmentation models versus the baseline, for example Morfessor 2.0 (Virpioja *et al.*, 2013).

To train the baseline morphological segmenter, we used Morfessor 2.0 toolkit [10] with Python interpreter. The training, development and testing datasets for Morfessor are the same as the datasets used to train our proposed segmenter for both Inuktitut and Inuinnaqtun, respectively. We filtered out all tokens of the corpus which are not included in the corresponding word list. These smaller datasets were also used in the semi-supervised training experiments. The F1 scores converged after 5 iterations for all runs. As the evaluation metric, we used the micro-average segmentation boundary F1-score. The scores were calculated based on the word types in the testing sets.

To train our NMT model, we first used the *subword-nmt* (Sennrich *et al.*, 2016) toolkit to create a $30k$ BPE joint source-target vocabulary. Then, to train our Transformer-based NMT models, we used the *Marian-nmt* toolkit (Junczys-Dowmunt *et al.*, 2018) with the following hyperparameter settings: 6-layer depth for both encoder and decoder, 8-layer multi-heads, embedding dimension of 512, hidden layers of 2,048 units in the feed-forward networks, with optimizer Adam and an initial learning rate of 0.0003. For the architecture type, we could choose either the Seq2Seq (Sutskever *et al.*, 2014) or the Transformer (Vaswani *et al.*, 2017) inside the toolkit. We performed multiple NMT experiments as follows:

– System 1 (*Baseline*): We chose the same configuration as described in Joanis *et al.* (2020), with only BPE-preprocessed data;

– System 1 + align information: We incorporated source-target word alignment information in the training step. We applied an unsupervised word aligner, *fast_align* (Dyer *et al.*, 2013) to generate symmetrized source-target alignments, trained on BPE preprocessed data;

– System 1 + lex.s2t: We combined the source-target bilingual lexicon, during the decoding phase, in the baseline system. We applied the lexicon extractor from *Moses* (Koehn *et al.*, 2007) to prepare a bilingual lexical shortlist;

– System 1 + align information + lex.s2t: We combined both word alignment information and the source-target bilingual lexicon in the baseline system;

– Systems 2, 3, 4, 5: We configured the proposed morphological segmentation using the standard or scholar-seeded settings combined with the sequence-to-

10. Morfessor 2.0 toolkit: `https://morfessor.readthedocs.io/en/latest/index.html`.

sequence based or the Transformer-based architectures for our NMT model, named AG-Standard+s2s, AG-Scholar+s2s, AG-Standard+TF, AG-Scholar+TF, respectively;

– Multilingual NMT system (multiNMT): We performed the following experiments applying the word segmentation for the source-side indigenous languages, *e.g.* Inuktitut, Inuinnaqtun, within different multilingual NMT systems:

- (multiNMT) We chose, for this baseline, the same configuration as described in Joanis *et al.* (2020), with only BPE-preprocessed data, with all source-target language pairs and the test set on Inuktitut-English only,

- (multiNMT-1) The training datasets are without segmenting any indigenous language (Inuktitut, Inuinnaqtun),

- (multiNMT-2) The source-side training datasets are segmented only for Inuktitut but not for Inuinnaqtun,

- (multiNMT-3) The source-side training datasets are segmented for both Inuktitut and Inuinnaqtun,

- (multiNMT-4) The source-side training datasets are segmented for Inuinnaqtun but not for Inuktitut.

4.3. *Evaluations and discussion*

4.3.1. *Morphological segmentation task*

We evaluated the morphological segmentation system using the automatic metrics: *Precision (P)*, *Recall (R)*, and *F1 score*.

For the supervised morphological segmentation, we evaluated only the Inuktitut data source. As described in Table 8, our proposed system, with all pretrained embeddings, showed a good performance with 75.33% in terms of F1 score. However, the Morfessor system outperformed our proposed system, with a gain of +4.37 points in terms of F1 score. Using the additional golden annotated data with the training data, the Morfessor model obtained better precision and recall than our proposed model, with a gain of +1.36 points and +6.83 points. In addition, the Morfessor model used an n-best Viterbi algorithm that allows extraction of all posible segmentations for a compound and the probabilities of the segmentation.

	Precision	**Recall**	**F1 score**
Morfessor	82.15	77.40	**79.70**
supervised_Inuktitut_WS	80.79	70.57	75.33

Table 8. *Results for Inuktitut supervised morphological segmentation.*

For the unsupervised morphological segmentation, we evaluated both Inuktitut and Inuinnaqtun data sources. Table 9 shows the performance and results of our models versus Morfessor for the polysynthetic language on the Inuktitut test set. The standard setting is better than the baseline, with a gain of +8.30 points in terms of precision,

on the test set, compared with Morfessor. Moreover, we also observed large gains of +8.92 points in terms of precision, on the test set, when using the scholar-seeded setting compared with Morfessor. All models obtained low recall between 77.40% and 82.33%, including Morfessor, due to the under-segmentation.

Inuktitut	Precision	Recall	F1 score
Morfessor	82.15	77.40	79.70
AG-Standard	90.45	81.51	85.75
AG-Scholar	**91.07**	**82.33**	**86.48**

Table 9. *Morphological segmentation task: Results for the Inuktitut test set using the Standard setting (AG-Standard), Scholar seeded setting (AG-Scholar), and Morfessor toolkit. Values in bold refer to the best performances.*

Table 10 shows the performance results of our models versus Morfessor for the polysynthetic language using the Inuinnaqtun test set. Both AG-based models outperformed the baseline, with gains of +3.33%, +17.62% in terms of F1 score, for the AG-standard setting and AG-Scholar setting, respectively. The recall of all three models is good enough to recognize all possible patterns, with 75.40%, 80.30%, and 82.83% for the baseline, AG-standard setting and AG-Scholar setting, respectively. However, the baseline and the AG-Standard model obtained low precision with 48.29% and 50.76%, respectively, compared with the AG-scholar-seeded model, which obtained 71.06%. We observed an under-segmentation in these models.

Inuinnaqtun	Precision	Recall	F1 score
Morfessor	48.29	75.40	58.87
AG-Standard	50.76	80.30	62.20
AG-Scholar	**71.06**	**82.83**	**76.49**

Table 10. *Morphological segmentation task: Results for the Inuinnaqtun test set using the Standard setting (AG-Standard), Scholar seeded setting (AG-Scholar), and Morfessor toolkit. Values in bold refer to the best performances.*

We noted that our proposed models could not correctly recognize more complex morphemes due to the languages' linguistic irregularities and rich morphophonemics. In particular, they were unable to detect common affixes such as *ag*, *ik*, *iq*, *mi*, *ti* or *ut* in Inuktitut and common lexical suffixes such as *at*, *aq*, *iq*, *na*, *ng* or grammatical ending suffixes such as *a*, *k*, *q*, *t*, *n*, *it*, *mi* or *uk* in Inuinnaqtun.

Tables 11 and 12 illustrate some predictions by all the models and the performance of our models on Inuktitut and Inuinnaqtun, respectively.

4.3.2. *Machine translation task*

We conducted additional evaluations of Machine Translation task based on the BLEU scores (Papineni *et al.*, 2002) which were computed with lowercase and *v13a* tokenization, using *sacrebleu* (Post, 2018). We also used chrF++ (Popović, 2015) to

Segmentation	Sentence Example
Raw text	niqtunaqtuq piku tusaattialaurit
Reference	niqtu naq tuq piku tusaa tti ala u rit
Uqailaut analyzer (Farley, 2012)	niqtu naq tuq piku tusaa ttia lau rit
BPE (Sennrich *et al.*, 2016)	niqtunaqtuq piku tusa@@ atti@@ alaur@@ it
Our proposed approach	niqtu naq tuq piku tusaa tti ala u rit

Table 11. *Illustrations of the Inuktitut word segmentation (Meaning: Mr. Picco, please listen).*

Word	Ground Truth	Morfessor	AG-Standard	AG-Scholar
aullarnanga	aullar na nga	aulla rn anga	aulla rna nga	aullar na nga
aullaqtinnatin	aullaq tinna tin	aulla q ti nna t in	aulla q tinna tin	aullaq tinna tin
igluptun	iglu ptun	iglu p tun	iglu ptun	iglu ptun
nattiqhiuqtuq	nattiq hiuq tuq	nattiq hi uq tu q	nattiq hi uq tuq	nattiq hiuq tuq
kitkungnin	kitku ngnin	kitku ng nin	kitku ng ni n	kitku ngnin
tupaktuhi	tupak tu hi	tupa k tu h i	tupak tu hi	tupak tu hi
iqaluktinnagu	iqaluk tinna gu	iqaluk ti nna gu	iqaluk tin na gu	iqaluk tin na gu

Table 12. *Illustrations of Inuinnaqtun morpheme segmentation predictions on the test set using the different settings such as Standard (AG-Standard), Scholar seeded (AG-Scholar), and Morfessor. Red text indicates deviations in segmentation from the Ground Truth.*

calculate the F1-score averaged on character n-gram precision and recall enhanced with word n-grams for the translation references and their hypotheses.

Experiment	BLEU (dev set)	BLEU (test set)	chrF++
System 1 (*Baseline*)	41.40	35.00	65.40
System 1 + align information	41.45	35.71	65.59
System 1 + lex.s2t	41.66	35.93	65.97
System 1 + align information + lex.s2t	41.78	**36.03**	66.30

Table 13. *Performance on Inuktitut-English NMT in terms of lowercase word BLEU score, using only the BPE subword segmentation method.*

We observed that combining the word alignment information and the source-target bilingual lexicon had a positive impact on the performance of the NMT model. Compared to the baseline, with all the additional features, the NMT system obtained a gain of +1.03 points in terms of BLEU score (Table 13). However, using only the BPE subword segmentation method, the multilingual NMT system outperformed the system 1 and all variants, with a gain of +3.06 points in terms of BLEU score (Table 14).

To go further, we performed multiple variants of the multilingual NMT systems, with and without applying our proposed word segmenter to the source-side indigenous languages (Table 14). We noted that the multiNMT-1 system obtained the worst

Experiment	BLEU (test set)	chrF++
multiNMT (baseline, only BPE method)	**38.06**	**68.15**
multiNMT-1 (-INU_segmented, -IKT_segmented)	8.11	14.52
multiNMT-2 (+INU_segmented, -IKT_segmented)	40.91	73.25
multiNMT-3 (+INU_segmented, +IKT_segmented)	**41.40**	**74.13**
multiNMT-4 (-INU_segmented, +IKT_segmented)	38.08	68.19

Table 14. *Performance of all the multilingual NMT for Inuktitut-English, with and without applying our proposed segmenter, where INU and IKT refers to Inuktitut and Inuinnaqtun, respectively.*

performance, only 8.11% in terms of BLEU score or 14.52% in terms of chrF++, due to without any indigenous languages. The best performance was obtained by the multiNMT-3, 41.40% in terms of BLEU score or 74.13% in terms of chrF++. We observed that the translation quality were significantly improved as we segmented indigenous languages in the source side rather than other related languages, up to +3.34 points versus the multiNMT baseline (Table 14). It means that is sufficient to segment the source-side indigenous languages to have a better performance. The related languages are not necessarily required for word segmentation.

Moreover, we tested other NMT systems with our proposed morphological segmentation based on Adaptor Grammars. All our systems 2, 3, 4 and 5 outperformed the baseline with gains of up to +2.98 points and +3.41 points in terms of BLEU score, on the development set and the test set, respectively (Table 15), compared to the baseline.

Experiment	dev	test	chrF++
System 1-*Baseline*-(Joanis *et al.*, 2020)	41.40	35.00	65.40
System 2 (AG-Standard+s2s)	43.93	37.78	66.43
System 3 (AG-Scholar+s2s)	**44.38**	**38.41**	**68.71**
System 4 (AG-Standard+TF)	44.18	38.28	68.41
System 5 (AG-Scholar+TF)	44.41	38.32	68.61

Table 15. *Performance on Inuktitut-English NMT in terms of lowercase word BLEU score, with our proposed morpheme segmenter.*

We compared our best system against other NMT systems from WMT 2020 using morphological segmentation methods, such as Roest *et al.* (2020), and Knowles *et al.* (2020). Our best system outperformed all the NMT systems from WMT 2020 in terms of BLEU score on the third version of the Nunavut Hansard test set, with 38.41% versus 30.05% and 29.90% (Table 16).

Roest *et al.* (2020) reported their best NMT system results due to multiple reasons. First, they applied three methods of segmentation: unsupervised such as BPE, LMVR (Ataman *et al.*, 2017) and 3-step segmentation. They varied the value of the decoder's penalty length based on results on the development set with 0.8 for news and 1.4

Experiment	dev	test
System (Knowles *et al.*, 2020)	–	29.90
System (Roest *et al.*, 2020)	–	30.05
Our best system 3 (AG-Scholar+s2s)	**44.38**	**38.41**

Table 16. *Comparison of performance results of our best system on Inuktitut-English NMT in terms of lowercase word BLEU score with other best systems of WMT 2020.*

System	Sentence Example
Raw	apiqqutiga turaaqtittumajara aanniaqarnangittulirijikkut ministangannut.
Reference	I would like to direct my question to the Minister of Health.
Baseline	This is a question for the Minister of Health.
System 1	My question is for the Minister responsible for Health.
System 2	My question is directed for the Minister of Health.
System 3	I would like to direct my question to the Minister of Health.
System 4	My question is directed for the Minister of Health.
System 5	I would like to ask my question for the Minister of Health.
MultiNMT	This is my question directed for the Minister responsible for Health.

Table 17. *Illustrations of some translation predictions using different NMT systems, from Inuktitut to English.*

for Hansards, respectively. Furthermore. there was a mixture of in-domain and out-domain training data. Finally, the use of ensembling and fine-tuning on all NMT systems helped to improve the BLEU performance.

In the case of Knowles *et al.* (2020), the final systems were trained on a mix of news and Hansard data, using joint BPE, BPE-dropout, tagged back-translation for Inuktitut-English, fine-tuning, ensembling, and the use of domain-specific models.

We assume the preprocessing as word segmentation helped to solve the complex morphology of Inuktitut at source-side. Our proposed NMT model outperformed the state-of-the-art, as presented in Joanis *et al.* (2020), using only BPE-preprocessed training data, with the best performance of 44.38% and 38.41% in terms of BLEU on the development set and the test set, respectively.

5. Conclusion and Perspectives

In this paper, we empirically explored different word segmentation techniques on both Inuktitut and Inuinnaqtun. We then proposed a novel morphological segmentation technique that may be applied to any indigenous language.

In the supervised approach, the neural networks-based word segmentation model showed promising results, but not good enough, due to multiple factors, such as the quantity of the annotated data and the quality of the pretrained embedding models.

In the semi-supervised and non-supervised approaches, the Adaptor Grammars-based word segmentation models yielded better results, employing a collection of grammatical rules from grammar books, and a lexicon from relatively little data. We applied our word segmenter to preprocess the Inuktitut source-side language before implementing an Inuktitut-English NMT system. We empirically evaluated our proposed NMT method against several baseline NMT architectures. Our proposed NMT system outperformed the state-of-the-art, as described in Joanis *et al.* (2020), with just BPE-preprocessed training data.

Our study makes an important contribution by focusing on morpheme segmentation in the source-side indigenous language. This significantly enhances the MT performance in the low-resource scenario. Furthermore, the NLP community is becoming increasingly interested in indigenous languages. Indigenous language research might lead to a more thorough knowledge of human languages and the development of universal NLP models.

In the future, we plan to add more annotated data and study other domain-specific characteristics to increase the segmentation model's accuracy. Moreover, we are developing a multilingual NMT framework in order to include more indigenous languages, particularly endangered ones, with the goal of preserving and revitalizing endangered and indigenous languages, as well as their legacy and culture. Globally, our research interest focuses on an inclusive, fairer and more equitable and responsible Artificial Intelligence, while emphasizing on the revitalization and preservation of indigenous languages. We have encountered a variety of challenges, including language skills, data gathering, and validation, to name a few. Thus, we seek to conduct research "by and with" indigenous peoples, which will help validate the results and construct more reliable linguistic resources that will be, we hope, of great help to the indigenous communities.

6. References

Aharoni R., Johnson M., Firat O., "Massively multilingual neural machine translation", *arXiv preprint arXiv:1903.00089*, 2019.

Arppe A., Schmirler K., Harrigan A. G., Wolvengrey A., "A Morphosyntactically Tagged Corpus for Plains Cree", *49th Algonquian Conference, Montreal, Quebec*, p. 27-29, 2017.

Ataman D., Negri M., Turchi M., Federico M., "Linguistically motivated vocabulary reduction for neural machine translation from Turkish to English", *arXiv preprint arXiv:1707.09879*, 2017.

Botha J. A., Blunsom P., "Adaptor Grammars for Learning Non- Concatenative Morphology", Association for Computational Linguistics, 2013.

Carpenter B., "Coding chunkers as taggers: Io, bio, bmewo, and bmewo+", *LingPipe Blog*p. 14, 2009.

Creutz M., Lagus K., "Unsupervised models for morpheme segmentation and morphology learning", *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 4, n^{o} 1, p. 1-34, 2007.

Dyer C., Chahuneau V., Smith N. A., "A simple, fast, and effective reparameterization of ibm model 2", *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 644-648, 2013.

Eskander R., Klavans J. L., Muresan S., "Unsupervised Morphological Segmentation for Low-Resource Polysynthetic Languages", *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, p. 189-195, 2019.

Eskander R., Rambow O., Muresan S., "Automatically tailoring unsupervised morphological segmentation to the language", *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, p. 78-83, 2018.

Farley B., "The uqailaut project", *URL http://www.inuktitutcomputing.ca*, 2012.

Gasser M., "Computational morphology and the teaching of indigenous languages", *Indigenous Languages of Latin America Actas del Primer Simposio sobre Enseñanza de Lenguas Indígenas de América Latina*, p. 52, 2011.

Godard P., Besacier L., Yvon F., Adda-Decker M., Adda G., Maynard H., Rialland A., "Adaptor grammars for the linguist: Word segmentation experiments for very low-resource languages", *Workshop on Computational Research in Phonetics, Phonology, and Morphology*, p. 32-42, 2018.

Haddow B., Bawden R., Barone A. V. M., Helcl J., Birch A., "Survey of Low-Resource Machine Translation", *arXiv preprint arXiv:2109.00486*, 2021.

Hochreiter S., Schmidhuber J., "Long Short-Term Memory", *Neural Comput.*, vol. 9, nº 8, p. 1735-1780, November, 1997.

Joanis E., Knowles R., Kuhn R., Larkin S., Littell P., Lo C.-k., Stewart D., Micher J., "The Nunavut Hansard Inuktitut English Parallel Corpus 3.0 with Preliminary Machine Translation Results", *Proceedings of The 12th Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, p. 2562-2572, May, 2020.

Johnson H., Martin J., "Unsupervised learning of morphology for English and Inuktitut", *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2*, Association for Computational Linguistics, p. 43-45, 2003.

Johnson M., "Unsupervised word segmentation for Sesotho using adaptor grammars", *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, p. 20-27, 2008.

Joshi P., Santy S., Budhiraja A., Bali K., Choudhury M., "The State and Fate of Linguistic Diversity and Inclusion in the NLP World", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, p. 6282-6293, July, 2020.

Junczys-Dowmunt M., Grundkiewicz R., Dwojak T., Hoang H., Heafield K., Neckermann T., Seide F., Germann U., Aji A. F., Bogoychev N. *et al.*, "Marian: Fast neural machine translation in C++", *arXiv preprint arXiv:1804.00344*, 2018.

Kann K., Mager M., Meza-Ruiz I., Schütze H., "Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages", *arXiv preprint arXiv:1804.06024*, 2018.

Knowles R., Stewart D., Larkin S., Littell P., "NRC Systems for the 2020 Inuktitut-English News Translation Task", *Proceedings of the Fifth Conference on Machine Translation*, p. 156-170, 2020.

Koehn P., Hoang H., Birch A., Callison-Burch C., Federico M., Bertoldi N., Cowan B., Shen W., Moran C., Zens R. *et al.*, "Moses: Open source toolkit for statistical machine translation", *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, p. 177-180, 2007.

Kudlak E., Compton R., *Kangiryuarmiut Inuinnaqtun Uqauhiitaa Numiktitirutait — Kangiryuarmiut Inuinnaqtun Dictionary*, vol. 1, Nunavut Arctic College: Iqaluit, Nunavut, 2018.

Le T. N., Sadat F., "Revitalization of Indigenous Languages through Pre-processing and Neural Machine Translation: The case of Inuktitut", *Proceedings of the 28th International Conference on Computational Linguistics*, p. 4661-4666, 2020.

Littell P., Kazantseva A., Kuhn R., Pine A., Arppe A., Cox C., Junker M.-O., "Indigenous language technologies in Canada: Assessment, challenges, and successes", *Proceedings of the 27th International Conference on Computational Linguistics*, p. 2620-2632, 2018.

Lowe R., *Basic Siglit Inuvialuit Eskimo Grammar*, vol. 6, Inuvik, NWT: Committee for Original Peoples Entitlement, 1985.

Mager M., Gutierrez-Vasques X., Sierra G., Meza-Ruiz I., "Challenges of language technologies for the indigenous languages of the Americas", *Proceedings of the 27th International Conference on Computational Linguistics*, Association for Computational Linguistics, Santa Fe, New Mexico, USA, p. 55-69, August, 2018.

Mager M., Oncevay A., Ebrahimi A., Ortega J., Gonzales A. R., Fan A., Gutierrez-Vasques X., Chiruzzo L., Lugo G. G., Ramos R. *et al.*, "Findings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas", *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, p. 202-217, 2021.

Micher J., "Improving coverage of an inuktitut morphological analyzer using a segmental recurrent neural network", *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, p. 101-106, 2017.

Micher J., "Using the Nunavut Hansard data for experiments in morphological analysis and machine translation", *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, p. 65-72, 2018.

Mithun M., "Morphological complexity and language contact in languages indigenous to North America", *Linguistic Discovery*, vol. 13, n° 2, p. 37-59, 2015.

Nicholson J., Cohn T., Baldwin T., "Evaluating a morphological analyser of Inuktitut", *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, p. 372-376, 2012.

Nicolai G., Coates E., Zhang M., Silfverberg M., "Expanding the JHU Bible Corpus for Machine Translation of the Indigenous Languages of North America", *Proceedings of the Workshop on Computational Methods for Endangered Languages*, vol. 1, p. 1-5, 2021.

Papineni K., Roukos S., Ward T., Zhu W.-J., "BLEU: a method for automatic evaluation of machine translation", *Proceedings of the 40th annual meeting on association for computational linguistics*, Association for Computational Linguistics, p. 311-318, 2002.

Pitman J., Yor M., "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator", The Annals of Probability – JSTOR, p. 855-900, 1997.

Popović M., "chrF: character n-gram F-score for automatic MT evaluation", *Proceedings of the Tenth Workshop on Statistical Machine Translation*, p. 392-395, 2015.

Post M., "A Call for Clarity in Reporting BLEU Scores", *Proceedings of the Third Conference on Machine Translation: Research Papers*, Association for Computational Linguistics, Brussels, Belgium, p. 186-191, October, 2018.

Provilkov I., Emelianenko D., Voita E., "Bpe-dropout: Simple and effective subword regularization", *arXiv preprint arXiv:1910.13267*, 2019.

Rice K., "Documentary linguistics and community relations", *Language Documentation & Conservation*, vol. 5, p. 187-207, 2011.

Roest C., Edman L., Minnema G., Kelly K., Spenader J., Toral A., "Machine Translation for English–Inuktitut with Segmentation, Data Acquisition and Pre-Training", *Proceedings of the Fifth Conference on Machine Translation*, p. 274-281, 2020.

Ruokolainen T., Kohonen O., Virpioja S., Kurimo M., "Supervised morphological segmentation in a low-resource learning setting using conditional random fields", *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, p. 29-37, 2013.

Schwartz L., Tyers F., Levin L., Kirov C., Littell P., Lo C.-k., Prud'hommeaux E., Park H. H., Steimel K., Knowles R. *et al.*, "Neural polysynthetic language modelling", *arXiv preprint arXiv:2005.05477*, 2020.

Sennrich R., Haddow B., Birch A., "Neural Machine Translation of Rare Words with Subword Units", *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Berlin, Germany, p. 1715-1725, August, 2016.

Sirts K., Goldwater S., "Minimally-supervised morphological segmentation using adaptor grammars", *Transactions of the Association for Computational Linguistics*, vol. 1, p. 255-266, 2013.

Sutskever I., Vinyals O., Le Q. V., "Sequence to sequence learning with neural networks", *Advances in neural information processing systems*, p. 3104-3112, 2014.

Tiedemann J., "Parallel data, tools and interfaces in OPUS.", *Lrec*, vol. 2012, Citeseer, p. 2214-2218, 2012.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I., "Attention is all you need", *Advances in neural information processing systems*, p. 5998-6008, 2017.

Virpioja S., Smit P., Grönroos S.-A., Kurimo M. *et al.*, "Morfessor 2.0: Python implementation and extensions for Morfessor Baseline", 2013.

Wang L., Cao Z., Xia Y., De Melo G., "Morphological segmentation with window LSTM neural networks", *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Yang J., Zhang Y., Dong F., "Neural Word Segmentation with Rich Pretraining", *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Vancouver, Canada, p. 839-849, July, 2017.

Notes de lecture

Rubrique préparée par Denis Maurel

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

Dimitrios MELETIS. The Nature of Writing : A Theory of Grapholinguistics. *Fluxus Editions* (coll. « Grapholinguistics and its applications », vol. 3). 2020. IX, 461 pages. ISBN : 978-2-9570549-2-3.

Lu par **Pascal Vaillant**

Université Sorbonne Paris Nord (Paris 13), LIMICS

Il est très difficile de définir les concepts d'une linguistique de l'écrit en s'abstrayant des spécificités d'une langue ou même d'une écriture, tant celles-ci diffèrent dans leurs unités et leurs principes combinatoires. L'ouvrage The Nature of Writing : A Theory of Grapholinguistics, *de Dimitrios Meletis, est la tentative la plus aboutie dans cette direction.*

Dimitrios Meletis, universitaire autrichien, livre avec *The Nature of Writing* une somme théorique impressionnante sur la langue écrite. L'ambition de Meletis est tout à la fois de définir un cadre sémiotique général de l'écriture (partie *Description*) et de fournir des guides d'application permettant d'analyser, dans ce cadre, la manière dont les utilisateurs des langues font usage des systèmes d'écriture et la manière dont ces derniers « s'adaptent » aux langues qu'elles représentent (partie *Explanation*). Il s'y est appliqué avec une rigueur impressionnante et a creusé en profondeur chaque concept et chaque outil méthodologique, jusqu'à obtenir un système théorique cohérent.

Définition du cadre

La première partie de l'ouvrage s'applique à fournir un cadre descriptif unifié à la linguistique de l'écrit (l'auteur utilise le terme *grapholinguistics*, calqué de l'allemand *Schriftlinguistik*[1]), qui puisse ensuite être appliqué à l'analyse de plusieurs langues écrites en utilisant des concepts communs. L'écrit est ici défini comme lié à une langue : les tentatives de sémasiographie pure (comme le *Bliss*) n'entrent pas dans le champ d'étude. Il n'est pas pour autant une simple transposition de la langue orale et « possède une vie propre » : en témoigne la

1 Dimitrios Meletis est issu du monde académique germanophone, où s'est développée depuis quatre décennies une tradition forte de linguistique de l'écrit.

présence, dans presque toutes les écritures, d'unités (les signes de ponctuation) ou d'oppositions (entre majuscules et minuscules) n'ayant pas d'équivalent dans la langue parlée.

L'auteur a voulu, dans cette première partie, élaborer des concepts génériques à partir d'une masse de travaux antérieurs qui ne formaient pas un tableau cohérent. Il ne l'a pas fait en réinventant la roue : l'abondance de ses sources montre une très grande culture du domaine, qu'il fait partager au lecteur en en dressant l'état de l'art. Il a en revanche constaté que les travaux antérieurs ne convergeaient pas vers des définitions superposables de concepts aussi fondamentaux que ceux de graphème ou d'unité élémentaire de l'écrit (le premier n'étant d'ailleurs pas, dans tous les travaux, défini comme la seconde). Si les réponses à ces questions sont différentes, c'est souvent que les questions n'ont pas été formulées de la même manière, ni avec les mêmes présupposés : alors que certains ont cherché à définir la structure d'une écriture comme un système de signes graphiques et sont partis des formes pour étudier leur combinatoire, puis leurs fonctions, d'autres l'ont abordée avant tout comme un système de transcription d'une langue orale et sont partis d'une typologie des unités linguistiques pour étudier les différentes manières de les représenter. Ces deux approches conduisent à des unités parfois complètement différentes. En outre, les travaux sur l'écriture, à part les ouvrages à caractère historique comme ceux de Marcel Cohen ou d'Ignace J. Gelb, sont rarement informés de la diversité des écritures du monde, comme l'auteur le note en préambule : « *descriptions of writing systems coexist but rarely reference one another* ». Or la conception que l'on peut se faire du fonctionnement de l'écrit diffère largement selon que l'on décrit l'écriture du chinois ou celle de l'allemand. L'impressionnante culture linguistique de Meletis en a fait l'un des rares à pouvoir réfléchir en profondeur aux questions interlinguistiques sans se contenter de généralités ou de ouï-dire, comme on en voit souvent (notamment concernant le chinois). Le premier apport de l'ouvrage est donc une cartographie du domaine étudié, la langue écrite, qui catégorise trois grands modules au sein desquels les objets d'étude sont différents : la graphétique, la graphématique et l'orthographe.

La *graphétique* décrit la combinatoire interne et externe des unités d'un système d'écriture indépendamment de la langue (dans ce sens restreint, *script*). Dans toutes les écritures, les segments de base (qu'il s'agisse de lettres ou de caractères) se distinguent les uns des autres par la disposition de formants graphiques plus fondamentaux. La combinatoire de ces formants, dans l'espace « segmental », est l'un des paramètres d'une typologie des écritures. C'est également dans le cadre de la graphétique que se définit l'allographie, qui est la possibilité pour un même élément abstrait (la lettre |a| par exemple) d'apparaître sous plusieurs formes (A, a, ɑ)[2]. Au-delà de l'espace segmental, la graphétique caractérise également la manière dont les unités se combinent dans l'espace de la ligne, puis dans l'espace de la page.

2 Certaines de ces variations sont régulées par le plan graphématique (comme les variantes positionnelles des lettres arabes ou l'opposition majuscules minuscules) ; d'autres sont « libres » ou plus exactement associées à des valeurs connotatives (comme le choix de la police de caractères). Dans tous les cas leur description relève du module graphétique.

La graphétique est le lieu de la description du matériau graphique de l'écriture ; cependant l'analogie entre graphétique et phonétique (suggérée par leur suffixe commun) a ses limites. En effet, si la phonétique est la description de la pure substance de la langue parlée (la forme lui étant imposée de l'extérieur, par des oppositions qui se manifestent dans les signifiants d'une langue donnée), la graphétique possède en elle-même une opposition forme/substance : que l'alphabet latin serve à noter l'allemand ou le wolof, les différences dans l'espace graphique qui permettent de distinguer la lettre E de la lettre F sont les mêmes.

La *graphématique* établit le lien entre les formes utilisées pour l'écriture et le système linguistique qui lui sert de support. Le concept central de la graphématique est le graphème et l'un des apports majeurs de Meletis est d'en proposer une définition à la fois rigoureuse et généralisable.

Beaucoup de travaux antérieurs ont eu tendance soit à éviter le terme de graphème, soit à l'utiliser pour désigner un élément d'un système d'écriture (en quelque sorte comme un mot à consonance savante pour « lettre »), soit à en proposer une définition *ad hoc* liée à un usage particulier (par exemple pour désigner les segments de chaîne écrite correspondant à des phonèmes, usage fréquent dans le domaine de la pédagogie de l'orthographe). Nina Catach, que l'auteur cite peu (peu de ses sources sont en français) proposait en 1979 une définition générique du graphème (« plus petite unité de la chaîne écrite ayant un correspondant phonique et/ou sémique susceptible d'une analyse linguistique »), mais tout le développement ultérieur de son propos utilisait le terme pour établir une typologie des fonctions des lettres et séquences de lettres en français. Semblablement, comme le note l'auteur, beaucoup de définitions sont spécifiques à une langue et à un système d'écriture. Meletis propose donc de caractériser le graphème par trois critères : (1) c'est un segment distinctif au niveau du sens (distinctif, mais pas nécessairement *uniquement* distinctif, puisqu'il peut correspondre à un morphème, comme c'est le cas général dans le chinois écrit) ; (2) il correspond à une unité fonctionnelle du système linguistique (mais, point important, pas nécessairement une unité du niveau phonologique) ; (3) il doit être minimal (ce qui exclut par exemple |q| de l'inventaire des graphèmes de l'allemand, puisqu'il n'y a pas une paire minimale où |q| apparaît sans |u|) ; le critère de minimalité n'implique pas qu'il s'agisse d'un *segment* minimal dans le découpage linéaire de la chaîne écrite, ce qui permet de rendre compte, par exemple, des jamo coréens. La robustesse de cette définition est éprouvée par son instanciation fructueuse dans toutes les grandes familles de systèmes d'écriture : allemand, thaï, chinois, japonais, coréen.

Au-delà du graphème, ce plan est également celui où s'élaborent les concepts de mot graphématique et de phrase graphématique.

L'*orthographe*, enfin, est le module qui décrit les contraintes d'usage, culturellement et socialement normées, qui s'imposent aux combinaisons d'unités graphématiques. Elles décrivent un sous-ensemble de l'espace des combinaisons possibles.

Évaluer l'adaptation des écritures et des orthographes

La deuxième partie du livre de Meletis est celle où se met en évidence l'utilité du cadre théorique proposé, y compris pour les lecteurs qui ne sont pas fascinés par les classifications *per se*. L'auteur met en effet ses concepts à l'épreuve en les utilisant pour expliquer comment les utilisateurs des langues écrites adaptent les systèmes d'écriture à leurs besoins. Le cadre conceptuel de la théorie de la naturalité (expliqué dans le premier chapitre de cette partie) lui sert à définir des outils méthodologiques permettant d'évaluer différents niveaux d'« adaptation » (*fit*) des écritures aux langues qu'elles représentent. Cette adaptation est mesurée structurellement, au niveau du système linguistique, mais également sur le plan de la performance, ainsi que dans la sphère sociolinguistique (valeurs socialement attribuées à l'écriture et aux différentes façons d'écrire).

Cette partie permet d'éclairer certains débats. Il est, par exemple, notoirement habituel, dans l'espace francophone, de se diviser sur les mérites de l'orthographe. Elle est décrite par certains comme « illogique » (par opposition à des orthographes jugées plus « transparentes », comme celle de l'espagnol). Meletis aide à considérer cette question plus globalement en ajoutant au critère de transparence phonologique celui de transparence morphématique.

Une meilleure compréhension des choix de représentation formelle

L'apport de l'ouvrage pour la linguistique informatique est indirect, du fait que la plupart des questions liées aux particularités de la langue écrite ont déjà, en 2021, trouvé des solutions par défaut. Ainsi, la diversité des unités linguistiques, de leurs niveaux de définition et de leurs règles de combinaisons, a été soigneusement prise en charge par la norme Unicode (qui est bien plus qu'un répertoire de formes graphiques) et les trois décennies de travail de spécialistes qui l'ont produite. Le livre de Meletis permet plutôt de mieux comprendre et d'éclairer rétrospectivement ce travail et d'en apprécier l'ingéniosité (pour ne citer qu'un exemple, dans la représentation du coréen).

La définition des mots, qui passe par une étape graphétique (pour les langues alphabétiques, par exemple, l'extraction des segments séparés par des espaces), puis par une étape graphématique (détermination des cas où certains signes comme l'apostrophe ou le trait d'union font partie du mot, repérage des mots composés, des verbes séparables en allemand…) est en général un choix par défaut intégrés dans l'étape de *tokenization* des chaînes de TAL et que la communauté fait sans y penser, à l'exception notable des travaux sur les langues asiatiques. L'ouvrage de Meletis a le mérite d'en éclairer les mécanismes sous-jacents.

Conclusion

Le livre de Meletis est une somme d'érudition et une impressionnante nouvelle classification cohérente des connaissances sur l'écrit, dans un cadre commun, adaptable à toutes les langues écrites. Et c'est une classification qui n'ignore pas les travaux antérieurs, mais les incorpore et s'en informe. Mais l'intérêt de cet ouvrage ne se limite pas à une belle ontologie de la linguistique de l'écrit. Il fournit également des outils conceptuels pratiques, précieux pour faire avancer, boussole en

main, des discussions qui avaient souvent tendance à rester indécidables, faute d'être ancrées à un socle théorique solide et à des indicateurs bien définis.

L'ouvrage *The Nature of Writing* s'impose d'emblée comme la référence bibliographique majeure de tous les travaux futurs de linguistique de l'écrit.

Vivi NASTASE, Stan SZPAKOWICZ, Preslav NAKOV, Diarmuid Ó SÉAGDHA. Semantic Relations Between Nominals, Second Edition. *Morgan & Claypool publishers*. 2021. 218 pages. ISBN : 9-781-63639-088-8.

Lu par **Yannis HARALAMBOUS**

IMT Atlantique, UMR CNRS 6285 Lab-STICC

Vivi Nastase a soutenu sa thèse à Ottawa, auprès de l'illustre Rada Mihalcea – les deux chercheuses ont un point commun : leur alma mater*, qui est l'université technique de Cluj-Napoca en Roumanie. En 2003 se forme le noyau des auteurs du présent ouvrage, quand Nastase publie avec Stan Szpakowicz (également de l'université d'Ottawa), un article qui porte déjà sur les relations sémantiques. En 2007, Preslav Nakov (de Berkeley) ainsi que trois autres personnes se joignent à eux pour proposer un défi lors de la conférence SemEval, défi qui porte sur le même sujet que cet ouvrage : la classification de relations sémantiques entre nominaux. En 2013, Diarmuid Ó Séaghdga (de Cambridge, UK) se joint à eux et ils transforment les résultats du défi, ainsi que l'expérience accumulée au fil des années, en volume de la collection Synthesis Lectures de Morgan & Claypool. La deuxième édition de celui-ci est l'ouvrage que nous nous proposons de décrire. Sorti en 2021, il a presque doublé de taille (on est passé de 120 à 218 pages). Mis à part quelques ajouts mineurs aux quatre chapitres de la première édition, la véritable contribution de la deuxième édition consiste en l'arrivée d'un cinquième chapitre sur – tendance actuelle oblige – le* deep learning*.*

L'ouvrage comporte cinq chapitres, le premier n'étant qu'une brève mise au point terminologique ainsi qu'une mise en garde sur ce que le livre « n'est pas ». À noter qu'on y trouve une définition du nominal : il ne s'agit pas d'un groupe nominal comme on pourrait le croire mais, plus spécifiquement, d'un nom, d'un nom propre, d'un nom déverbal, d'un nom désadjectival, d'un nom précédé d'un modificateur (l'exemple donné est celui d'un participe passé suivi d'un nom : *processed food*) ou, de manière récursive, d'une suite de nominaux. En effet, là où le français utilise plutôt des groupes prépositionnels, l'anglais va accumuler des noms qui se suivent sans aucune préposition, ainsi, la « couleur de feuille de chêne » sera en anglais une suite de quatre noms : *oak tree leaf color*.

Relations entre nominaux, relations entre concepts

Comme indiqué dans le titre du chapitre 2, l'ouvrage emprunte deux points de vue en parallèle : celui du traitement automatique de la langue (auquel cas on s'intéresse aux nominaux dans les textes) et celui de la représentation des connaissances (auquel cas on s'intéresse aux relations entre concepts, ce qui constitue l'armature même des ontologies, les arêtes de leur structure de graphe).

Après un aperçu historique des interactions entre langue et connaissance, les auteurs enchaînent sur un très intéressant historique des relations sémantiques, d'abord dans les textes (et donc entre nominaux), et ensuite dans les ontologies (et donc entre concepts). Dans la première partie (les textes), il est passionnant de parcourir les tentatives de classification de relations sémantiques, entre 1826 (par l'un des frères Grimm !) et aujourd'hui. Dans la deuxième partie (les ontologies), on s'aperçoit que les problématiques sont les mêmes, même si ceux qui ont proposé des classifications s'intéressent plutôt aux propriétés des relations (du type : faut-il distinguer les méronymies transitives des intransitives ?) qu'à celles des arguments.

Pour mettre un peu d'ordre dans tout cela, les auteurs de l'ouvrage proposent une liste des dimensions de variation des relations sémantiques : l'aspect ontologique ou idiosyncratique (autrement dit : la relation dépend-elle du contexte ?), l'arité de la relation (les auteurs ratent une formidable occasion de parler des graphes conceptuels de Chein & Mugnier), la question de l'ouverture ou fermeture de l'ensemble des relations, l'ordre des relations (second ordre = relations de relations), le domaine et le seuil de précision des relations.

Ce chapitre est vraiment très intéressant pour qui veut avoir une vision globale de la modélisation des relations sémantiques, surtout lorsque l'on se pose des questions générales sur les méthodes et les outils à utiliser, selon le matériau textuel disponible et les objectifs à atteindre.

Extraction supervisée de relations sémantiques

Au chapitre 3 se pose la question de l'extraction *supervisée* de relations sémantiques, dans le sens où les classes de relations sont connues à l'avance et où, idéalement, on a déjà annoté les arguments de relations potentielles. Il commence par un historique des nombreux défis qui ont été posés dans des conférences : MUC (*Message Understanding Conference*, 1987-1997), ACE (*Automatic Content Extraction*, 1999-2008) et Sem-Eval 2007. Ensuite, après une section dédiée aux travaux spécifiques aux arguments de type nom-nom, il est question de certains corpus collaboratifs comme la Wikipédia (et en particulier les info-boîtes), DBpedia, YAGO, WikiNet et Freebase (qui, entre-temps, a été racheté par Google et ensuite injecté dans WikiData), ainsi que de corpus dans des domaines spécialisés, comme la médecine. Après cette partie introductive, on entre dans le vif du sujet : les propriétés (que les datamineurs appellent *features*) et les algorithmes d'apprentissage, où l'on retrouve toutes les méthodes classiques.

Ce chapitre est très bien écrit et très synthétique, il fournit un bon panorama des méthodes classiques (d'avant le *deep learning*) avec énormément de références.

Extraction non supervisée ou semi-supervisée de relations sémantiques

Vu l'extrême variété de relations possibles et l'énorme quantité de données textuelles dont nous disposons aujourd'hui, l'extraction non supervisée – et donc ouverte à tout – est bien plus prometteuse en matière de résultats que l'extraction supervisée. Le chapitre retrace un grand nombre d'approches, à commencer par les approches historiques : extraction de relations à partir d'un dictionnaire en 1981, utilisation de motifs pour extraire la relation d'hyponymie par Hearst en 1992. Ce

travail est très intéressant parce qu'il ne s'est pas limité au cas direct (« X est Y ») mais a exploité des cas indirects comme « X, de même que Y », « X, y compris Y », etc. : dans les deux cas on peut déduire que « Y est un X »). L'approche de Hearst a servi à initier un *bootstrapping* : on trouve des relations, on s'en sert pour former des nouveaux motifs (à trous), et on recommence. Ce faisant, le nombre de relations augmente de manière significative, mais aussi le bruit provenant de mauvaises interprétations, cette source d'erreur est appelée *dérive sémantique* et des mesures de spécificité et de confiance ont été définies pour y remédier. Plus tard, d'autres ont utilisé le *clustering* pour trouver des classes de termes et en déduire des relations sémantiques, ainsi que des motifs comme ceux de Hearst, mais cette fois-ci pour réunir les termes qui participent à une conjonction dans un même cluster (par exemple, dans « X tels que Y_1 et Y_2 », où Y_1 et Y\down{2} sont cohyponymes de X).

Relations sémantiques et *deep learning*

Ce chapitre, qui est le principal ajout de la seconde édition de l'ouvrage, nous fait parcourir la quasi-totalité des techniques du *deep learning*, depuis les réseaux de neurones récurrents jusqu'aux transformeurs (BERT et compagnie), en passant par les réseaux convolutifs, les LSTM et biLSTM et le mécanisme d'attention. La présentation est bien structurée et ne présuppose qu'un minimum de connaissances dans le domaine. En plus, elle se place au juste milieu entre les présentations orientées code, que l'on trouve dans des ouvrages qui poussent comme des champignons depuis quelques années, et les présentations mathématiques avec des notations à vous faire dresser les cheveux sur la tête et que l'on trouve dans des ouvrages d'obédience plutôt mathématique.

Après deux sections courtes et purement introductives, la troisième section du chapitre s'intéresse à la modélisation des attributs à travers les plongements de mots, en commençant par la création de plongements à partir de textes. Il y est question d'abord des deux méthodes désormais classiques (*skip-gram* et sac de mots continu), de la possibilité d'utiliser des espaces métriques autres que l'espace euclidien pour les plongements et de l'introduction de données concernant le contexte des mots dans les plongements à travers les BiLSTM et les différents BERT. Ensuite, on s'intéresse à la création de plongements de mots (ou d'entités) à partir de structures ontologiques. *A priori,* cela est plus simple puisque tout est déjà structuré, mais on est aussi plus exigeant : plus qu'un simple contexte, on se propose de capter *toute* la structure de graphe de l'ontologie. Les problématiques afférentes sont discutées et accompagnées de références à de nombreux travaux dans le domaine.

Après avoir couvert toutes les facettes de la modélisation des attributs, on passe à la modélisation des relations sémantiques. Si, dans le paragraphe précédent on a parlé de « mots » et non pas de « nominaux », c'était pour faciliter l'approche par les réseaux de neurones. Mais ici on n'y échappe plus : une relation sémantique est forcément plus complexe qu'un « mot » et une première partie de cette section s'intéresse aux différentes manières de représenter des structures complexes (phrases, chemins dans un arbre syntaxique, chemins hiérarchiques dans WordNet, etc.) en tant que données d'entrée de réseau de neurones. Dans la

deuxième partie, il y a un véritable saut qualitatif : plutôt que d'essayer par tous les moyens de « linéariser » les structures complexes qui nous intéressent, on garde la structure de graphe et on utilise des réseaux de neurones de graphes, les travaux dans cette direction ne manquent pas à l'appel.

La section qui suit s'intéresse aux données. Après un court descriptif des corpus disponibles, on revient à la méthode, déjà décrite au quatrième chapitre, de supervision distante, méthode qui fournit des corpus très bruités. Pour remédier à ce problème de bruit, trois méthodes sont décrites : les réseaux adversariaux génératifs (où un générateur se bat contre un discriminateur pour séparer le grain de l'ivraie), les réseaux avec mécanisme d'attention et l'apprentissage à renforcement qui fonctionne comme un jeu, avec des états, des actions et des récompenses.

Arrivé·à ce point, on réalise que les sections 1 à 5 de ce chapitre n'ont fait que poser les fondements nécessaires à la section 6 qui est la plus longue et la plus dense du chapitre. De la même manière que l'on a considéré la modélisation de mots dans la section 3, il s'agit ici (enfin) de modélisation apprentissage de relations sémantiques. On considère tout à tour l'apprentissage de relations à partir de structures ontologiques, à partir de textes et à partir des deux réunis. Toutes les méthodes possibles et imaginables y passent, et rien que de les énumérer ferait sauter la limite de pages allouée aux notes de lecture !

Conclusion

Certains ouvrages de la collection *Synthesis Lectures on Human Language Technologies* sont focalisés sur des sujets assez pointus et, de ce fait, n'intéressent qu'un public restreint. Ce n'est pas le cas de celui-ci. Que ce soit par la culture débordante de ses auteurs ou par l'universalité de son sujet (on retrouve les relations sémantiques dans *tous* les domaines du TAL), cet ouvrage est un trésor d'informations et de connaissances et ne manquera pas d'enrichir son lectorat, qui peut être de tout niveau (du doctorant au chercheur confirmé). En plus, il se lit agréablement et fait preuve d'un bon équilibre entre érudition et technicité, entre grands principes et petits conseils pratiques. Nous le conseillons vivement et attendons avec impatience la troisième édition, qui sortira sans doute dans les années 2030-2035, et dans laquelle le *deep learning* sera relégué aux vestiges du passé, au profit de nouvelles technologies dont nous ignorons encore tout aujourd'hui.

Anne ABEILLÉ, Danièle GODARD. La Grande Grammaire du Français. *Éditions Actes Sud – Imprimerie nationale*. 2021. 2 628 pages. ISBN 978-2-330-14239-1.

Lu par **Dominique LEGALLOIS**

Université Paris 3 – Sorbonne nouvelle / LaTTiCe

Une année après le très bel ouvrage « Cette histoire de la phrase française » *(octobre 2020), les éditions Actes Sud font paraître la très attendue* Grande Grammaire du Français *dirigée*

par Anne Abeillé et Danièle Godard, avec la collaboration d'Annie Delaveau et Antoine Gautier. On rappellera que le projet a été initié en 2002 sous l'égide du CNRS et de la DGLFLF (délégation générale à la langue française et aux langues de France). Il a reçu le soutien de nombreux laboratoires de recherche et universités françaises ou étrangères. Une soixantaine de linguistes ont participé à la rédaction des deux volumes (plus de 2 500 pages) qui s'intègrent dans un coffret cartonné vert foncé. L'objet est en soi très beau. Il existe également une version Internet par abonnement, qui donne notamment la possibilité d'écouter des extraits sonores, car cette Grande Grammaire *accorde une place importante à la langue parlée.*

Encyclopédique, la *Grande Grammaire du Français* (désormais GGF) s'adresse à un grand public averti, mais aussi aux étudiants, enseignants et bien sûr aux linguistes. Son dispositif permet de multiples adresses : des zones grisées, souvent en début de section, présentent des éléments fondamentaux ; des descriptions plus fouillées ou plus techniques, à destination des lecteurs plus avertis, sont présentées en retrait. Des tableaux très utiles récapitulent des formes catégorisées, illustrées d'exemples. Cinquante fiches synthétiques sont proposées à la fin de l'ouvrage : elles décrivent la nature et la fonction de mots grammaticaux (*même, pour, où*, etc.) ou bien des fonctionnements grammaticaux (accord du verbe, de l'adjectif, inversion du sujet) ; leur utilité pour la préparation aux concours d'enseignement est évidente. Un tableau (en introduction) fait la synthèse de quelques-uns des changements terminologiques les plus importants, entre la terminologie de 1998 et celle de 2020. Un précieux glossaire de trente pages donne la définition des catégories employées. *Cinquante-six pages de références concluent l'ouvrage, mais la fin de chaque chapitre mentionne des repères bibliographiques en lien direct avec les questions traitées.*

L'une des caractéristiques principales de la GGF est qu'elle a été réalisée à partir des outils de la linguistique moderne : les corpus et les bases textuelles (par exemple, le Corpus de français parlé au Québec, les corpus Valibel, Frantext, le corpus « 88milSMS », Clapi, etc.) ont permis de calculer des fréquences définissant des usages (du standard et du non-standard) ou bien des variations (diatopiques, de registre) selon l'origine des données. En cela, la GGF se veut fondamentalement descriptive : non pas par souci de non-normativité, mais par souci de décrire la langue française telle qu'elle se parle et s'écrit, dans sa diversité même, depuis les années cinquante, dans une partie de la francophonie. En effet, les grammaires françaises « sur le marché », dont personne ne peut raisonnablement dire qu'elles sont normatives, se fondent encore sur l'écrit littéraire, mais surtout, en réalité, sur des exemples inventés appropriés. Si la GGF illustre également le plus souvent ses rubriques par des exemples de ce type (quatre mille exemples attestés jalonnent l'ouvrage), elle les invente en fonction des données empiriques recueillies et analysées. En cela, ces exemples reflètent l'usage et leur manipulation vient justifier un type d'analyse approprié. Ils sont de plus annotés, concernant leur acceptabilité, selon diverses catégories : inacceptable (*), inacceptable dans le contexte (#) (*Jean est arrivé, mais je ne sais pas où*), douteux (?), variable selon les locuteurs (%) (*il a mangé ici plusieurs personnes*), non conforme à la norme (*tu peux les faire manger la viande* – Louisiane) (!). La GGF n'est cependant pas une grammaire de l'usage, au sens strict du terme.

La GGF se compose de vingt chapitres, avec une structuration à la fois classique et particulière. Heureusement classique, car on ne voit pas comment une grammaire pourrait échapper aux catégories fondamentales de la phrase, du verbe, du syntagme nominal, etc. Particulière, parce que certains chapitres sont consacrés exclusivement à des catégories sémantiques (*la négation ; le temps, l'aspect, le mode*) et aussi parce que d'autres chapitres développent des notions que l'on s'attendrait à voir traitées dans une partie plus englobante, par exemple le chapitre « *La coordination et la juxtaposition* » reprend et complète celui sur « *La phrase* » (chapitre 1). « *Le verbe* » introduit beaucoup d'éléments repris et développés dans le chapitre suivant, « *Les constructions verbales fusionnées* ». Le titre de ce chapitre atteste, par ailleurs, certaines nouveautés terminologiques (plus habituellement on parlerait de *complexe verbal* plutôt que de *fusion*). Celui intitulé « *Les proformes* » traite de la présentation d'une catégorie plus générale que celle traditionnelle des pronoms (ainsi y figurent également toute forme dont l'interprétation nécessite le contexte ou la situation : les déterminants démonstratifs et possessifs, les adverbes interrogatifs (*quand*), les adjectifs *tel* et *quel*, etc.). Contrairement aux autres types de subordonnées, la subordonnée complétive ne fait pas l'objet d'un chapitre – non plus d'ailleurs que la subordonnée participiale, quelque peu ressuscitée sans trop que l'on sache pourquoi. Les quatre derniers chapitres sont des chapitres d'interface particulièrement bien documentés (cf. la forme sonore des énoncés, l'ancrage des énoncés dans l'énonciation, les écritures numériques…).

Parmi les choix innovants de la GGF, on note par exemple la notion de « marqueur », qui vient préciser la nature prépositionnelle de *de* et *à* employés devant l'infinitif (*Paul continue de / à travailler*) et qui se justifie par le fait que la préposition peut être séparée de l'infinitif par *ne pas*, *ne plus* – elle est alors considérée comme introducteur de syntagme verbal (comme le subordonnant (conjonction) *que* est marqueur, car il introduit un syntagme phrastique). *Ici*, *là*, *là-bas*, *dehors*, *partout*, etc. traditionnellement considérés comme des adverbes, sont, dans la GGF, classés parmi les prépositions sans complément en raison de la difficulté à les employer entre l'auxiliaire et le participe passé (ou devant l'infinitif) : **Paul est dehors allé* ; **je ne veux pas ici aller*. Pour la même raison, *où* perd son statut d'adverbe et même de relatif « plein » : il est une préposition soit interrogative, soit relative (et relative sans antécédent), soit concessive (*où que tu ailles*). Remarquons tout de même que *où* est donné comme adverbe de lieu dans le chapitre consacré aux proformes : cette petite incohérence montre la difficulté qu'ont dû rencontrer les directrices de la GGF pour homogénéiser chez les soixante contributeurs, la terminologie et les analyses. La fonction *coordonné* s'applique à tous les membres d'une coordination ; la fonction *extrait* aux membres disloqués ou aux thèmes suspendus.

Autres spécificités : *certains* (dans *certains m'ont fasciné*) ou encore *les miens*, *les tiens*, etc. sont analysés comme des syntagmes nominaux sans nom, et non comme des pronoms. Ainsi *certains* garde dans l'exemple sa nature de déterminant devant un nom en ellipse. Il s'agit là justement de proformes. En revanche, *ce dernier* est bien un pronom, car il ne peut être modifié. La notion d'*intransitif* est étendue aux verbes à complément oblique (*j'ai rêvé de vous*) : l'impotent transitif

indirect est jeté aux orties – qu'il y reste ! On parle dans la GGF d'*omission du complément* plutôt que d'emploi absolu du verbe. Les notions de voix ou de diathèse sont rangées dans un vieux tiroir : *alternances de valence* suffira. Le complément direct perd son objet, ce qui permet d'éviter la distinction entre COD et complément essentiel. *Phrase désidérative* remplace *phrase à l'impératif*. Le terme *proposition*, en fait assez peu employé, est restreint à un emploi particulier : il désigne le contenu sémantique d'une phrase déclarative et non plus une configuration formelle.

On peut bien sûr faire quelques critiques, nécessairement éparses ici. C'est d'ailleurs le propre d'une grammaire que de se soumettre à la question du linguiste. Ainsi, la GGF entérine la fonction d'*ajout* à la place de l'encombrant *complément circonstanciel* ; cependant, la catégorie de l'ajout concerne des cas qui nous semblent relativement différents puisque cette fonction s'applique non seulement aux éléments circonstants, mais également aux adjectifs considérés parfois en emploi attribut (*elle est partie furieuse*), aux incises (*Lou, je crois, a terminé son travail*), aux pronoms contrastifs (*Paul viendra, lui*) ou quantifieurs (*les élèves viendront tous*). La fonction *ajout* est donc très générale, ce que reconnaissent explicitement les autrices. De même, on regrette aussi que la fonction *oblique* (plus universelle que celle de *complément indirect*) soit trop englobante : elle ne permet pas d'établir de distinction entre certains compléments indirects et les compléments datifs (*Paul parle de Marie à Marc*), alors que les pronominalisations de ces compléments sont pourtant bien différentes (cette remarque vaut en fait pour presque toutes les grammaires du français). On peut aussi se demander si la multiplication des fonctions n'ajoute pas une certaine complexité descriptive : si le linguiste peut y trouver son compte, le pédagogue, lui, a du souci à se faire… Sauf erreur de notre part (car la GGF est un océan vaste à naviguer), la particularité sémantique et syntaxique des verbes labiles (*Paul cuit les pâtes* vs *les pâtes cuisent*), qui aurait pu illustrer l'alternance de valence, ne fait pas l'objet d'une description, ni même d'une mention, alors que ces verbes (verbes ergatifs selon certaines terminologies) représentent un nombre important de lexèmes verbaux parmi les plus courants. Ces quelques remarques – qui pourraient, elles aussi, être discutées – peuvent se justifier globalement par la difficulté à embrasser et apprécier en peu de temps, un volume tout à fait considérable d'aspects descriptifs nouveaux : ce qui est fondamentalement présent dans la GGF est une logique descriptive et analytique dont l'appropriation ne peut être que progressive.

En résumé : la GGF est une somme considérable qui fait déjà date dans l'histoire de la grammaire du français. C'est un ouvrage de référence dont l'aspect encyclopédique et l'approche « usage » viennent combler une lacune dans la description du français. Il s'agit là d'un magnifique outil dont on n'a pas fini de manipuler les pages.

Résumés de thèses et HDR

Rubrique préparée par Sylvain Pogodalla

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

Maria BORITCHEV : maria_boritchev@yahoo.fr

Titre : Modélisation dynamique des dialogues

Mots-clés : sémantique formelle, linguistique computationnelle, dialogue, questions.

Title: *Dialogue Modeling in a Dynamic Framework*

Keywords: *formal semantics, computational linguistics, dialogue, questions.*

Thèse de doctorat en informatique, LORIA, UMR 7503, Université de Lorraine, Nancy, sous la direction de Maxime Amblard (MC HDR, Université de Lorraine, LORIA) et Philippe de Groote (DR, INRIA, LORIA). Thèse soutenue le 22/11/2021.

Jury : M. Maxime Amblard (MC HDR, Université de Lorraine, LORIA, codirecteur), M. Philippe de Groote (DR, INRIA, LORIA, codirecteur), M. Miguel Couceiro (Pr, Université de Lorraine, LORIA, président), Mme Farah Benamara Zitoune (MC HDR, Université Paul Sabatier, IRIT, rapporteuse), M. Jonathan Ginzburg (Pr, Université Paris Diderot – Paris 7, rapporteur), Mme Ellen Breitholtz (MC, University of Gothenburg, Suède, examinatrice), Mme Iris Eshkol-Taravella (Pr, Université Paris Nanterre, examinatrice).

Résumé : *L'étude formelle du discours soulève de nombreuses interrogations liées à la nature et à la définition des phrases et de la manière dont une suite de phrases s'articule pour former un discours cohérent. Le langage est intrinsèquement dynamique : dans sa sémantique en contexte (par exemple, l'utilisation de références) et dans l'interaction (par exemple, les liens entre les actes de dialogue). Le passage du discours au dialogue donne lieu à des questions plus spécifiques en particulier liées à la relation entre les questions et les réponses. Afin d'aborder ces thématiques, nous nous concentrons sur la sémantique des questions.*

Il existe de nombreux formalismes et cadres de travail pour la sémantique formelle des phrases déclaratives et du discours. Le dialogue, pour sa part, est largement étudié d'un point de vue linguistique et traitement automatique des langues. L'objectif de notre travail est d'utiliser les théories classiques de sémantique formelle dans un cadre orienté vers le dialogue réel. Cette thèse présente une formalisation sémantique du dialogue dans une théorie dynamique des types simples.

Nous produisons des modèles du dialogue, et en particulier de l'articulation des questions et des réponses, en mêlant la Neo-Davidsonian Event Semantics *à la* Inquisitive Semantics *de manière compositionnelle et dynamique à travers l'usage de la* Continuation Style Dynamic Semantics*. Notre modèle est ancré dans une implémentation d'interface syntaxe-sémantique appelée* Abstract Categorial Grammars*.*

Une autre façon d'aborder la sémantique du dialogue est de s'intéresser aux données réelles, ce qui permet de mettre en perspective nos idées formelles et de les confronter aux observations. Pour ce faire, nous avons constitué un corpus, appelé Dialogues in Games *(DinG), composé de transcriptions d'enregistrements de personnes jouant au jeu de société Catane (en français). Nos études, centrées sur les questions et les réponses dans des données orales multilingues (anglais, français, néerlandais, espagnol mexicain, italien du nord et mandarin), a donné lieu à plusieurs schémas d'annotation, dont une partie a été appliquée à DinG.*

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-03541628

Marie CANDITO : marie.candito@univ-paris-diderot.fr

Titre : Annoter et prédire des représentations linguistiques de phrases

Mots-clés : annotations linguistiques, analyse syntaxique automatique, analyse sémantique automatique.

Title: *Annotate and Predict Linguistic Representations of Sentences*

Keywords: *linguistic annotation, syntactic parsing, semantic parsing.*

Habilitation à diriger des recherches en informatique, Laboratoire de Linguistique Formelle, UMR 7110, Département de linguistique, Université de Paris. Habilitation soutenue le 19/01/2022.

Jury : M. Sylvain Schmitz (Pr, Université de Paris, président), M. Frédéric Béchet (Pr, Aix-Marseille Université, rapporteur), Mme Claire Gardent (DR, CNRS, LORIA, rapporteuse), Mme Paola Merlo (Pr, Université de Genève, Suisse, rapporteuse), M. Benoît Crabbé (Pr, Université de Paris, examinateur), M. Pierre Zweigenbaum (DR, CNRS, LISN, examinateur).

Résumé : *Le travail présenté, réalisé pour la majeure part en collaboration, concerne principalement l'explicitation de représentations linguistiques de phrases,*

qu'il s'agisse de la méthodologie de constitution manuelle de telles ressources, ou de la définition de modèles permettant de prédire de telles représentations, par apprentissage supervisé ou semi-supervisé.

Le mémoire présente, à divers degrés de détail :

– des contributions en termes de ressources annotées, pour le français, qu'il s'agisse d'expressions polylexicales, d'arbres de dépendances, de graphes de dépendances profondes, de cadres et rôles sémantiques FrameNet. Ces ressources sont définies avec exigence quant à la finesse des analyses linguistiques, et quant à leur utilisabilité comme données d'apprentissage supervisé ;

– des contributions en analyse syntaxique en dépendances, d'une part sur la problématique de la robustesse des analyseurs supervisés face aux mots inconnus et aux changements de domaine, d'autre part sur l'exploitation d'un contexte plus large et de modèles spécialisés pour la correction automatique d'arcs, pour les phénomènes le plus fréquemment source d'erreurs (rattachement prépositionnel et coordination) ;

– la proposition d'un modèle pour l'analyse automatique en graphes de dépendances reposant sur un apprentissage multitâche, où la tâche principale est réalisée par un parseur biaffine neuronal, et où des tâches auxiliaires sont définies pour ajouter de l'interdépendance dans la prédiction des arcs.

Ce mémoire couvre une période longue, marquée par l'arrivée de méthodes neuronales en TAL. L'apprentissage par transfert permet de fournir des représentations vectorielles de mots, hors ou en contexte, en utilisant des objectifs génériques, en particulier la prédiction d'un mot sachant son contexte. Il est fascinant de constater qu'un objectif aussi simple et brut permet de construire des modèles apportant des gains très importants dans à peu près toutes les tâches de TAL. Le transfert se fait en utilisant des corpus à l'état brut, ne nécessitant pas de modélisation linguistique (outre la définition des unités considérées). C'est ainsi l'objectif même d'analyse automatique de phrases qui est remis en cause. Certaines tâches, comme la traduction automatique, le résumé automatique, l'analyse de sentiments sont actuellement mieux gérées par des modèles « de bout-en-bout », ne nécessitant pas d'expliciter des représentations linguistiques traditionnelles. On assiste même à une ingénierie inversée, où ce sont les modèles de langue préentraînés sur corpus bruts qui sont sondés, pour voir si et où s'y cachent les concepts linguistiques traditionnels.

Cela dit, même s'il est difficile de prédire l'avenir du concept même d'analyse automatique de phrases, les besoins d'interprétabilité des modèles et de quantification des phénomènes linguistiques font que le concept reste d'actualité. On peut même espérer que les sondes linguistiques des modèles neuronaux permettent d'éclairer d'un jour nouveau certains concepts linguistiques.

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-03544267

Julie HUMBERT-DROZ : julie.humbertdroz@gmail.com

Titre : Définir la déterminologisation : approche outillée en corpus comparable dans le domaine de la physique des particules

Mots-clés : déterminologisation, linguistique outillée, linguistique de corpus, terminologie textuelle, langues de spécialité.

Title: *Defining Determinologisation: Tool-Based Approach in a Comparable Corpus in the Field of Particle Physics*

Keywords: *determinologisation, tool-based approach, corpus linguistics, textual terminology, languages for special purposes.*

Thèse de doctorat en sciences du langage & traitement informatique multilingue, Département de traitement informatique multilingue, Faculté de traduction et d'interprétation, Université de Genève, Suisse. Thèse soutenue le 07/09/2021.

Jury : Mme Aurélie Picton (Pr, Université de Genève, Suisse, codirectrice), Mme Anne Condamines (DR, CNRS, codirectrice), Mme Agnès Tutin (Pr, Université Grenoble Alpes, rapporteuse), M. Patrick Drouin (Pr, Université de Montréal, Canada, rapporteur), Mme Amélie Josselin-Leray (MC, Université Toulouse – Jean Jaurès, examinatrice), Mme Mathilde Fontanet (Pr, Université de Genève, Suisse, présidente).

Résumé : *Cette thèse porte sur la question de la déterminologisation dans le domaine de la physique des particules. La déterminologisation peut s'envisager à la fois comme un processus de passage de termes d'une langue de spécialité à la langue générale et comme le résultat de ce processus, c'est-à-dire le fonctionnement de termes dans la langue générale. Face au manque de travaux traitant spécifiquement le processus, notre travail vise à combler cette lacune en abordant cette question de manière systématique et approfondie. Nous nous intéressons particulièrement aux phénomènes sémantiques qui se produisent au cours du processus. Par la description de ces phénomènes, nous cherchons à mieux cerner la notion de déterminologisation et ses manifestations dans les textes.*

Dans ce but, nous avons développé une méthode de linguistique outillée pour l'analyse des termes dans des textes représentant différentes étapes du processus de déterminologisation. Cette méthode repose sur un corpus comparable et sur deux indices. Les données sont organisées en cinq sous-corpus, constitués de textes qui relèvent de différents genres textuels et degrés de spécialisation. Ces textes sont sélectionnés dans le but d'approcher le continuum entre langue de spécialité et langue générale. En outre, afin de caractériser les différents fonctionnements des termes dans le corpus, deux indices basés sur les contextes distributionnels des termes sont définis.

L'exploration de ces indices dans le corpus révèle la diversité des fonctionnements des termes dans la presse, en comparaison avec les textes spécialisés. Les phénomènes repérés permettent d'alimenter la réflexion sur la nature des changements sémantiques

résultant de la déterminologisation et montrent que ces changements reflètent des différences de points de vue ou de conceptualisation, inhérents à la multidimensionnalité des concepts. Les données mettent également en lumière les mécanismes à l'œuvre dans la création des emplois métaphoriques des termes qui se fondent non seulement sur les usages des termes dans la presse, lorsque de nouvelles connotations sont véhiculées, mais aussi sur la coexistence dans la langue générale des composants du terme ou d'unités de la même famille morphologique.

Nos observations contribuent ainsi à définir la déterminologisation comme un processus complexe et non linéaire de passage de termes dans la langue générale, qui fait intervenir de nombreux intermédiaires et qui est influencé par au moins cinq facteurs, aussi bien linguistiques qu'extralinguistiques. Plus largement, nous montrons l'ampleur des questions liées au processus d'intégration et au fonctionnement des termes dans la langue générale. En ce sens, notre thèse participe à reconnaître la transversalité de la déterminologisation dans les questions de circulation des termes, de diffusion des connaissances et de néologie.

URL où le mémoire peut être téléchargé :
https://archive-ouverte.unige.ch/unige:157351

Martin LENTSCHAT : martin.lentschat@gmail.com

Titre : Instanciation de relations n-aires dans des articles scientifiques guidée par une Ressource Termino-Ontologique de domaine

Mots-clés : relations n-aires, extraction d'information, extraction de relations, ingénierie des connaissances, ressource termino-ontologique, représentation de données, mesure de pertinence.

Title: *n-Ary Relations Instantiation from Scientific Articles Driven by a Domain Ontological and Terminological Resource*

Keywords: *n-ary relations, information extraction, relation extraction, knowledge engineering, ontological and terminological resource, data representation, relevance measure.*

Thèse de doctorat en informatique, UMR TETIS, Université de Montpellier, sous la direction de Patrice Buche (IR, INRAE, IATE, UMR 1208), Juliette Dibie (Pr, INRAE, MIA-Paris, UMR 518) et Mathieu Roche (DR, CIRAD, TETIS). Thèse soutenue le 14/12/2021.

Jury : M. Patrice Buche (IR, INRAE, IATE, UMR 1208, codirecteur), M. Patrice Bellot (Pr, Université Aix-Marseille, LIS, UMR 7020, rapporteur), Mme Nathalie Pernelle (Pr, Université Sorbonne Paris Nord, LIPN, UMR 7030, rapporteuse), Mme Nathalie Aussenac-Gilles (DR, CNRS, IRIT, UMR 5505, présidente), M. Konstantin Todorov (MC HDR, Université de Montpellier, LIRMM, UMR 5506, examinateur),

Mme Juliette Dibie (Pr, INRAE, MIA-Paris, UMR 518, codirectrice), M. Mathieu Roche (DR, CIRAD, TETIS, codirecteur).

Résumé : *Cette thèse s'inscrit dans le domaine de recherche des* smart data *et consiste à proposer de nouvelles méthodes de représentation et d'extraction de données expérimentales à partir d'articles scientifiques, évaluées sur un corpus dans le domaine des emballages alimentaires.*

L'objectif de cette thèse est de peupler une base de connaissances d'instances de relations n-aires extraites de documents scientifiques textuels. Les connaissances expérimentales visées sont représentées sous forme de relations n-aires composées d'arguments symboliques (un syntagme) et quantitatifs (une valeur numérique et une unité de mesure).

L'approche proposée s'appuie sur une ressource termino-ontologique (RTO) et se décompose en deux phases :

1) extraction des instances d'arguments,

2) mise en relation de ces instances dans des relations n-aires.

*La phase 1 propose une représentation des instances d'arguments extraites : SciPuRe (*Scientific Publication Representation*). Celle-ci intègre des descripteurs ontologiques, lexicaux et structurels. La phase 2 s'appuie sur les informations présentes dans les tableaux des documents, extraits automatiquement, pour guider l'extraction des relations n-aires à partir de relations partielles devant être complétées par les instances d'arguments de la phase 1. Trois approches sont proposées et évaluées afin d'identifier les instances d'arguments qui doivent compléter les relations : l'utilisation de la structure des documents, l'analyse des cooccurrences entre instances d'arguments, et l'utilisation de mesure de similarité donnée par des modèles de* word embeddings*.*

Nos résultats montrent l'importance du filtrage des instances à l'issue de la phase 1. Les deux critères mesurant la pertinence d'une instance d'argument symbolique étant sa spécificité et sa fréquence. La pertinence des arguments quantitatifs est déterminée par la discrimination de l'instance d'argument selon les sections des articles. Les résultats montrent un effet positif lors du filtrage de 20% des instances les moins pertinentes. En phase 2, nous avons expérimenté une approche d'assistance aux experts en sélectionnant plusieurs candidats pour chaque instance d'argument manquante dans une relation partielle. Nos résultats montrent que la méthode à adopter varie selon l'approche souhaitée. Lors de la sélection d'un seul candidat, l'approche fondée sur les analyses des cooccurrences donne les meilleurs résultats. Avec une sélection de trois ou cinq candidats, l'analyse des similarités sémantiques par des modèles BERT fournit de bons résultats. Enfin, lors de la sélection de dix candidats, l'approche fon-

dée sur la structure des documents est la plus efficace pour compléter les relations n-aires.

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-03587319

Didier SCHWAB : didier.schwab@univ-grenoble-alpes.fr

Titre : Contributions au Traitement Automatique des Langues et à un domaine d'application, la Communication Alternative et Augmentée

Mots-clés : traitement automatique des langues et de la parole, communication alternative et augmentée, clarification de sens, représentation du sens, acquisition du sens, exploitation du sens.

Title: *Contributions to Speech and Natural Language Processing and to an Application Domain, Alternative and Augmentative Communication*

Keywords: *speech and natural language processing, alternative and augmentative communication, meaning clarification, meaning representation, meaning acquisition, meaning exploitation.*

Habilitation à diriger des recherches en informatique, Laboratoire d'informatique de Grenoble, Université Grenoble Alpes. Habilitation soutenue le 08/12/2021.

Jury : M. Emmanuel Morin (Pr, Université de Nantes, rapporteur), M. Pierre Zweigenbaum (DR, CNRS, LISN, rapporteur), M. Andrei Popescu-Belis (Pr, Haute École spécialisée de Suisse occidentale, Vaud, Suisse, rapporteur), M. Mathieu Lafourcade (MC HDR, Université de Montpellier, examinateur), M. Laurent Besacier (scientifique principal, Naver Labs Europe, Grenoble, examinateur), Mme Pierrette Bouillon (Pr et doyenne, Université de Genève, Suisse, examinatrice), Mme Sophie Dupuy-Chessa (Pr, Université Grenoble Alpes, présidente).

Résumé : *Que feriez-vous si vous étiez privé de la parole, peut-être même incapable de faire le moindre geste ? Cette situation peut arriver à tout le monde de manière transitoire (voyage dans un pays, problème de santé) ou de manière permanente (handicap). Elle peut arriver simplement après une opération chirurgicale relativement lourde ou un accident. Dans une moindre mesure, elle peut également arriver si vous êtes dans un pays dont vous ne parlez pas la langue. On vous montrera des images, des pictogrammes pour que vous puissiez faire passer votre message, avoir besoin de quelque chose, chercher son chemin, indiquer un problème de santé. Une telle communication est appelée* communication alternative et augmentée *(CAA).*

Les recherches présentées ici se situent dans le domaine du traitement automatique des langues et de la parole (TALP). Elles concernent la représentation, l'acquisition et l'exploitation du sens pour et par la clarification des données langagières. L'exploration de la communication alternative et augmentée a rapidement conduit à se

demander comment les recherches menées au GETALP pourraient être bénéfiques à la CAA et, suivant un schéma de pensée classique, comment, en retour, le traitement automatique des langues et de la parole pourrait bénéficier de la CAA.

Ce document explique comment ces travaux, tout en restant axés sur le sens et sa clarification, sont passés petit à petit de l'écrit aux pictogrammes en intégrant la parole et le regard ; comment ils ont porté sur la désambiguïsation lexicale, les représentations vectorielles pour le TALP, la traduction automatique du texte et de la parole jusqu'à des travaux autour du dialogue. Enfin, il présente les hypothèses et pistes de recherches qui pourront être suivies dans les années qui viennent.

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-03535726