Journées GDR LIFT 2021

6-7 décembre 2021 – Grenoble

Comité d'organisation

Benjamin Lecouteux, LIG/Université Grenoble-Alpes
Maximin Coavoux, LIG/CNRS
Emmanuelle Esperança-Rodier, LIG/Université Grenoble-Alpes
Claire Lemaire, LIG/Université Grenoble-Alpes, LAIRDIL/Université Paul Sabatier
Timothée Bernard, LLF/Université de Paris
Patrick Caudal, LLF/Université de Paris

Comité de programme

Angélique Amelot
Pascal Amsili
Heather Burnett
Berthold Crysmann
Marco Dinarelli
Solène Evain
Karën Fort
Sylvain Kahane
Pierre Magistry
Alexis Michaud
Tatiana Nikitina
Carlos Ramish
Emmanuel Schang
Gilles Sérasset
Thierry Poibeau
Valentin Vydrine
Guillaume Wisniewski

Présentations invitées

Laurent Besacier (Naver Labs Europe)

Titre: Self-Supervised Learning for Low Resource Speech Tasks

Abstract: Self-supervised learning using huge unlabeled data has been successfully explored for image processing and natural language processing. Since 2019, recent works also investigated self-supervised representation learning from speech. They were notably successful to improve performance on downstream tasks such as speech recognition. These recent works suggest that it is possible to reduce dependence on labeled data for building speech systems through acoustic representation learning. In this talk I will present an overview of these recent approaches to self-supervised learning from speech and show my own investigations to use them in spoken language processing tasks for which size of training data is limited.

Alda Mari (Institut Jean Nicod, CNRS / ENS - EHESS - PSL)

Titre: Epistemic future in questions, MICA and evidence quality

Abstract: In our talk we address the question of the interpretation of Italian MICA in questions with future in Italian. Contrary to recent views according to which MICA is a common ground management operator, we propose that MICA is a metaevaluator that ranks possibilities as low in a normality scale and ultimately gives rise to an inference of surprise in questions with future. This analysis is cast in a framework where future is analyzed as an epistemic modal (Giannakidou and Mari 2018; in opposition to views according to which future is an evidential (Mari 2010, Eckardt and Beltrama 2019, Frana and Menendez-Benito 2019) and in which the modal skeleton contains a metaevaluator. We propose a distinction between evidence *source* and evidence *quality* that allows us to ground the epistemic use of future in degraded evidence leaving room for the uncertainty presupposition over which MICA operates.

Daan van Esch (Google Research)

Titre: Building Language Technology for Everyone

Matti Miestamo (University of Helsinki)

Titre: Negation: Language typology, documentation and description

Abstract: Cross-linguistic typological work on negation has paid most attention to standard negation, i.e. the negation of declarative verbal main clauses (Dahl 1979; Payne 1985; Dryer 2013a,b,c; Miestamo 2005, 2013). Other aspects of negation that have received at least some attention in large-scale typological studies include the negation of imperatives (van der Auwera and Lejeune 2013), the negation of stative (nonverbal, existential, etc.) predications (Croft 1991; Eriksen 2011; Veselinova 2013), the negation of indefinite pronouns (Haspelmath 1997, 2013; Van Alsenoy 2014), abessives (Stolz et al. 2007), the effects of negation on the marking of NPs (Miestamo 2014), and negative replies to questions (Holmberg 2015) – for an overview

of typological work on negation, see Miestamo 2017. Currently, typological work is underway on various aspects of the typology of negation: e.g., Veselinova's work on negative lexicalizations and the relationship between negation and TAM, Miestamo and Koptjevskaja Tamm's work on antonyms, Van Olmen's work on negative imperatives, and Mauri and Sansò's work on anticircumstantial clauses as well as Miestamo, Shagal and Silvennoinen's work on negation in dependent clauses. In this talk, I will give an overview of current typological knowledge of negation, focusing especially on interesting issues arising from my own work on the domain. I will also show how typological knowledge on negation can be made use of in language documentation and description, presenting a typologically and functionally oriented questionnaire for describing the domain of negation (Miestamo 2016) and giving an update on how the questionnaire is being used in an ongoing project describing negation in a number of languages from around the world.

References:

van der Auwera, J. and L. Lejeune. 2005. The prohibitive. In M. Haspelmath, M. Dryer, D. Gil and B. Comrie (eds.), The world atlas of language structures, 290–293. Oxford: OUP.

Croft, W. 1991. The evolution of negation. Journal of Linguistics 27(1). 1–27.

Dahl, Ö. 1979. Typology of sentence negation. Linguistics 17. 79–106.

Dryer, M. S. 2013a. Negative morphemes. In M. Dryer and M. Haspelmath (eds.), World atlas of language structures online. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/112/.

Dryer, M. 2013b. Order of negative morpheme and verb. In M. Dryer and M. Haspelmath (eds.), World atlas of language structures online. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/143.

Dryer, M. 2013c. Position of negative morpheme with respect to subject, object, and verb. In M. Dryer and M. Haspelmath (eds.), World atlas of language structures online. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/144.

Eriksen, P.K. 2011. 'To not be' or not 'to not be': The typology of negation of non-verbal predicates. Studies in Language 35 (2): 275-310.

Haspelmath, M. 1997. Indefinite pronouns. Oxford: Oxford University Press.

Haspelmath, M. 2013. Negative indefinite pronouns and predicate negation. In M. Dryer and M. Haspelmath (eds.), World atlas of language structures online. Leipzig: Max Planck Institute for Evolutionary Anthropology, http://wals.info/chapter/115.

Holmberg, A. 2015. The Syntax of Yes and No. Oxford: OUP.

Miestamo, M. 2005. Standard negation: The negation of declarative verbal main clauses in a typological perspective. Berlin: Mouton de Gruyter.

Miestamo, M. 2014. Partitives and negation: A cross-linguistic survey. In S. Luraghi and T. Huumo, eds., Partitive Cases and Related Categories, 63-86. Mouton de Gruyter.

Miestamo, Matti. 2016. Questionnaire for describing the negation system of a language. Available online via http://tulquest.huma-num.fr/fr/node/134.

Miestamo, M. 2017. Negation. In A. Aikhenvald and R. M. W. Dixon, eds., The Cambridge Handbook of Linguistic Typology, 405-439. Cambridge: CUP.

Payne, J. 1985. Negation. In T. Shopen (ed.), Language typology and syntactic description, volume I, Clause structure, 197–242. Cambridge: Cambridge University Press.

Stolz, T., C. Stroh and A. Urdze. 2007. WITH(OUT): On the markedness relations between comitatives/instrumentals and abessives. Word 58(2). 63-122.

Van Alsenoy, L. 2014. A New Typology of Indefinite Pronouns, with a Focus on Negative Indefinites. U Antwerp dissertation.

Veselinova, L. 2013. Negative existentials: A cross-linguistic study. Rivista di linguistica 25 (1): 107-145.

Table des matières

Analyse orientée corpus d'universaux de Greenberg sur Universal Dependencies

Hee-Soo Choi[1, 2] Bruno Guillaume[1] Karën Fort[1, 3]
(1) Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
(2) ATILF, Université de Lorraine & CNRS, F-54000 Nancy, France
(3) Sorbonne Université F-75006 Paris, France
hee-soo.choi@loria.fr, Bruno.Guillaume@loria.fr, karen.fort@loria.fr

MOTS-CLÉS : universaux linguistiques, typologie linguistique, ordre des mots, multilinguisme, Universal Dependencies, corpus arborés, GREW.

KEYWORDS: language universals, linguistic typology, word order, multilinguism, Universal Dependencies, treebanks, GREW.

1 Entre typologie linguistique et TAL

Avec l'expansion des données numériques, de plus en plus de ressources multilingues voient le jour et constituent des traces durables de nos pratiques langagières. Ces dernières sont particulièrement précieuses dans le domaine du Traitement Automatique des Langues (TAL) dont l'un des objectifs est de traiter toutes les langues naturelles. Dans nos travaux, nous nous sommes intéressés à la vérification d'universaux de Greenberg (Greenberg, 1966) de manière automatique et empirique sur les corpus d'Universal Dependencies (UD) (de Marneffe et al., 2021). L'étude des universaux repose fortement sur une tradition empirique et typologique mais également sur des connaissances tirées d'ouvrages de référence. À travers nos expériences, nous fournissons des résultats fondés uniquement sur de grandes quantités de données avec un échantillon de 141 corpus, soit 74 langues d'UD 2.7. Avec l'outil GREW (Guillaume, 2021), nous avons ainsi déterminé trois ordres de mots (l'ordre sujet - verbe - objet, l'ordre adposition - nom et l'ordre adjectif - nom) et vérifié quatre universaux. En faisant le choix de traiter chaque corpus individuellement, nous avons pu évaluer l'homogénéité entre corpus d'une même langue et analyser les raisons des possibles divergences. Enfin, notre étude sur 74 langues permet également de soulever des incohérences interlinguistiques liées au schéma d'annotations.

2 Contraintes méthodologiques

La vérification des universaux linguistiques ont demandé de faire des choix dans notre approche en raison de certaines contraintes. Tout d'abord, un travail de tri sur les universaux a dû être effectué. Les universaux de Greenberg, établis sur la base de 30 langues de différentes familles de langues, sont au nombre de 45 et sont relatifs à l'ordre des mots, la syntaxe et la morphologie. En raison des annotations d'UD qui ne nous permettent pas d'obtenir certaines informations notamment au niveau morphologique, nous avons choisi de traiter les universaux relatifs à l'ordre des mots. Concernant nos données, en mettant à disposition un schéma d'annotations universel, le projet UD s'inscrit dans

l'objectif de traiter un maximum de langues et de favoriser ainsi les recherches multilingues. Les corpus d'UD constituent donc des données à première vue optimales pour notre tâche de vérification d'universaux. Toutefois, cette tâche reste ambitieuse et est mise à mal par les différences notables qui résident entre les langues et qui se traduisent par des incohérences dans les annotations. Le caractère universel des annotations est donc à prendre avec précaution. Enfin, nous avons choisi de traiter 74 langues, qu'évidemment nous ne maîtrisons pas toutes. Cette contrainte nous impose donc de nous intéresser à des caractéristiques basiques et universelles afin d'éviter davantage de biais. Par ailleurs, nous faisons le choix de nous positionner au niveau du corpus et non au niveau de la langue, ce qui revient à traiter, en réalité, non pas 74 langues mais 141 corpus.

3 Des résultats en accord avec Greenberg et WALS

Avant d'entreprendre la vérification des universaux, nous avons déterminé trois ordres de mots pour les 141 corpus : i) l'ordre sujet - verbe - objet, ii) l'ordre adposition - nom et iii) l'ordre adjectif - nom. Nous avons utilisé GREW (Guillaume, 2021), un outil de réécriture de graphes permettant de faire des requêtes sur les corpus et d'extraire les occurrences d'une construction linguistique particulière.

Pour déterminer l'ordre dominant de manière quantitative, nous utilisons le même critère que WALS (Dryer and Haspelmath, 2013) qui considère un ordre comme dominant s'il présente une fréquence d'apparition au moins deux fois plus grande que le deuxième ordre le plus fréquent (Dryer, 2013). Nous avons donc calculé le ratio entre les deux ordres les plus fréquents. Si le ratio est supérieur ou égal à deux, on considère l'ordre le plus fréquent comme l'ordre dominant, sinon on considère que le corpus n'a pas d'ordre dominant (noté NDO pour *No Dominant Order*). Dans le cas où seulement deux ordres sont possibles (par exemple, adjectif - nom / nom - adjectif), si le ratio est supérieur à 2, la fréquence de l'ordre le plus fréquent est supérieure à $\frac{2}{3}$. Utiliser cette mesure nous a permis de comparer les résultats obtenus pour les trois ordres avec les données présentes dans WALS.

Nos résultats sont majoritairement en accord avec ceux de WALS pour les trois ordres, mis à part certaines exceptions pour lesquelles nous fournissons une analyse plus détaillée. Si les ordres adjectif - nom et adposition - nom présentent des résultats relativement tranchés avec une tendance marquée pour un des deux cas, l'ordre sujet - verbe - objet présente plus d'hétérogénéité entre les corpus d'une même langue. Pour cet ordre, nous avons soulevé plusieurs facteurs expliquant l'hétérogénéité : le genre des corpus (corpus oraux, corpus de romans, corpus de journaux, corpus de textes bibliques...), la période des corpus (pour les langues mortes notamment), les erreurs d'annotations et certaines spécificités des langues (topicalisation du sujet par exemple) (Choi et al., 2021).

Suite à la classification selon les trois ordres précédemment décrits, quatre universaux de Greenberg ont été vérifiés :

Universel 1 *Dans les phrases déclaratives avec un sujet nominal et un objet nominal, l'ordre dominant est presque toujours un ordre dans lequel le sujet précède l'objet.*

Dans nos résultats, sur 141 corpus, 91 sont SVO, 24 sont SOV, 4 sont VSO et 22 sont NDO, ce qui confirme l'universel de Greenberg pour 119 corpus et 59 langues. Concernant les corpus sans ordre dominant, si nous considérons uniquement l'ordre le plus fréquent sans calculer le ratio, tous les corpus présentent un des trois ordres où le sujet précède l'objet, à l'exception de deux corpus : l'`Amharic-ATT` et le `Latin-LLCT`.

Universel 3 *Les langues d'ordre dominant VSO sont toujours prépositionnelles.*

Universel 17 *Avec une fréquence largement supérieure à la normale, les langues d'ordre dominant VSO ont l'adjectif après le nom.*

Nous avons traité ces universaux ensemble dans la mesure où ils concernent tous deux les langues d'ordre dominant VSO. Le tableau 1 donne ainsi les corpus d'ordre dominant VSO avec leurs fréquences, ainsi que la proportion de prépositions et de l'ordre Nom - Adjectif. Les universaux sont vérifiés mais sur seulement quatre corpus.

Corpus	VSO	Prep	Nom - Adj
Arabic-NYUAD	54,56 %	99,97 %	99,69 %
Irish-IDT	99,14 %	99,78 %	98,91 %
Scottish_Gaelic-ARCOSG	97,49 %	100 %	84,82 %
Welsh-CCG	78,57 %	100 %	82,54 %

TABLE 1 – Proportions de prépositions et d'ordre Nom - Adjectif dans les corpus VSO.

Universel 4 *Avec une fréquence largement supérieure à la normale, les langues d'ordre normal SOV sont postpositionnelles.*

D'après nos résultats, 24 corpus sont d'ordre dominant SOV, ce qui correspond à 15 langues. Dix langues sont postpositionnelles : le bambara, le basque, le coréen, le hindi, le japonais, le kazakh, l'ouïghour, le ourdou, le telugu et le turc.

Les cinq langues restantes sont l'afrikaans, l'allemand, le latin, le perse et le sanskrit. L'afrikaans et le perse présentent des corpus fortement SOV mais sont prépositionnels. L'exception du perse est également soulignée par Greenberg bien que cette langue ne soit pas dans son échantillon. L'allemand et le latin sont des langues multicorpus mais elles ne comptent qu'un seul corpus d'ordre dominant SOV, leurs autres corpus étant considérés comme sans ordre dominant. Nous supposons que la formulation de Greenberg « ordre normal SOV » permet de ne pas prendre en compte ce type de langues dans son universel. Et enfin, pour le corpus du sanskrit, nous n'avons trouvé aucune occurrences de postposition ou préposition. En effet, le motif GREW utilisé détecte les postpositions et prépositions annotées avec l'étiquette ADP (adposition). Or le corpus du sanskrit présente l'étiquette PART (particule). Nous pouvons noter que nous retrouvons le même phénomène sur un des corpus du coréen, le `Korean-PUD`.

4 Les limites du schéma d'annotations universel

Si UD propose un noyau d'annotations universel, les annotateurs sont libres dans leurs choix d'annotations, ce qui provoque une certaine hétérogénéité entre les langues mais aussi entre corpus d'une même langue. Pour extraire les constructions linguistiques décrites précédemment, nous devions identifier précisément les annotations codant pour ces constructions. Lors de cette étape, nous avons relevé plusieurs incohérences, notamment un conflit entre deux étiquettes de partie du discours : ADP (adposition) et PART (particule). La distinction entre une particule et une adposition est délicate à définir, en particulier dans les langues agglutinantes. Les adpositions, les conjonctions de coordination et de subordination sont considérées comme des particules dans les directives d'UD, mais celles-ci demandent à privilégier l'étiquette la plus précise possible.

Par ailleurs, nous avons remarqué que certains corpus présentent un nombre conséquent de dépendants non-nominaux pour les relations `nsubj` et `obj`, ce qui est contradictoire avec les directives d'UD. Dans le même esprit, la relation `case` implique normalement des gouverneurs nominaux, ce qui n'est pas respecté par certains corpus.

Le schéma UD se base initialement sur un schéma d'annotations créé pour l'anglais, le Stanford Dependencies (de Marneffe et al., 2014), ce qui a orienté certains choix d'annotations. Pour les langues présentant une structure différente de l'anglais, adapter les annotations amène les créateurs des corpus à faire des choix d'annotations qui ne sont pas forcément en accord avec les autres corpus de la langue. Cela impacte la cohérence entre les langues ainsi qu'entre corpus d'une même langue.

5 Conclusion

Nos résultats confirment les observations linguistiques sur de grandes quantités de données. Notre travail peut ainsi compléter les bases de données typologiques comme WALS qui présente des brèches pour sept langues que nous avons traitées : l'afrikaans, le féroïen, le galicien, le kazakh, le maltais, le naija et le slovaque. En outre, nous nous sommes efforcés de fournir des analyses pour expliquer les incohérences détectées dans les annotations d'UD, soit en examinant les documentations relatives aux corpus, soit en faisant appel à des locuteurs natifs autour de nous. L'aspect collaboratif du projet nous a ensuite permis de faire des retours et ainsi contribuer à l'amélioration des corpus concernés.

Références

Choi, H.-S., Guillaume, B., Fort, K., and Perrier, G. (2021). Investigating Dominant Word Order on Universal Dependencies with Graph Rewriting. In *Recent Advances in Natural Language Processing (RANLP2021)*, en ligne, Bulgarie.

de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. (2014). Universal Stanford dependencies : A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4585–4592, Reykjavik, Islande. European Language Resources Association (ELRA).

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2) :255–308.

Dryer, M. S. (2013). Determining dominant word order. In Dryer, M. S. and Haspelmath, M., editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Dryer, M. S. and Haspelmath, M., editors (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Greenberg, J. H. (1966). Some universals of grammar with particular reference to the order of meaningful elements. In Greenberg, J. H., editor, *Universals of Human Language*, pages 73–113. MIT Press, Cambridge, Mass.

Guillaume, B. (2021). Graph matching and graph rewriting : GREW tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : System Demonstrations*, pages 168–175, Online. Association for Computational Linguistics.

Convertir le Trésor de la Langue Française en Ontolex-Lemon : un zeste de données liées

Sina Ahmadi[1*] Mathieu Constant[3]
Karën Fort[2,4] Bruno Guillaume[4] John P. McCrae[1]
(1) Insight Centre for Data Analytics, National University of Ireland Galway (2) Sorbonne Université
(3) ATILF, Université de Lorraine & CNRS (4) Université de Lorraine, CNRS, Inria, LORIA
sina.ahmadi@insight-centre.org, mathieu.constant@univ-lorraine.fr,
{karen.fort,bruno.guillaume}@loria.fr, john.mccrae@insight-centre.org

RÉSUMÉ

Nous présentons dans ce papier les travaux que nous avons réalisés pour convertir dans le modèle Ontolex-Lemon l'une des plus importantes ressources lexicographiques pour le français : le Trésor de la Langue Française. En effet, malgré l'utilisation généralisée de cette ressource, son format actuel, basé sur XML, ne respecte pas les standards les plus récents de la représentation des données lexicographiques, notamment ceux basés sur les données liées. Nos travaux mettent en lumière la nécessité d'établir des mécanismes permettant d'augmenter l'inter-opérabilité des ressources et des technologies pour créer et maintenir des ressources lexicographiques.

ABSTRACT

Revisiting the *Trésor de la Langue Française*

In this paper, we report our efforts to convert one of the most comprehensive lexicographic resources of French, the *Trésor de la Langue Française*, into the Ontolex-Lemon model. Despite the widespread usage of this resource, the original XML format seems to impede its integration in language technology tools. In order to breathe new life into this resource, we examine the usage and the conversion to more interoperable formats, primarily those based on the linguistic linked data, to provide this resource to a broader range of applications and users.

MOTS-CLÉS : Ressources lexicographiques, données linguistiques liées, TAL.

KEYWORDS: Lexical-semantic resources, linguistic linked data, natural language processing.

1 Introduction

Les ressources lexico-sémantiques sont des référentiels de connaissances qui présentent le vocabulaire d'une langue de manière descriptive, structurée ou conceptualisée. Parmi ces ressources, les dictionnaires sont les plus répandus et historiquement utilisés pour étudier les langues naturelles et les traiter grâce aux techniques de traitement automatique des langues (TAL) [1]. Par conséquent, ces dictionnaires jouent un rôle crucial dans plusieurs applications du TAL telles que la désambigüisation lexicale [2], l'étiquetage de rôles sémantiques [3] ou l'analyse syntaxique et lexicale [4]. Malgré le grand nombre de ressources issues des initiatives communautaires, comme le Wiktionnaire [1], les ressources créées par des experts restent fondamentales du fait de leur qualité et de leur degré d'élaboration.

∗. Le travail a été réalisé pendant une visite scientifique à l'ATILF.

1. https://fr.wiktionary.org

2 Ontolex-Lemon

Aux cours de ces dernières années, les standards basés sur les données liées et le Web sémantique ont changé l'éco-système de création, de représentation et de maintenance des ressources langagières, en particulier les dictionnaires [5, 6]. Les modèles de données tels qu'Ontolex-Lemon [7] définissent des ontologies en s'appuyant des ressources terminologiques et lexicales présentes sur le Web sémantique. Ces modèles permettent également d'augmenter l'inter-opérabilité et le multilinguisme des ressources en fournissant des mécanismes d'alignement de représentations sémantiques existantes sous forme d'ontologies [8, 9].

OntoLex-Lemon est un modèle basé sur LEMON – Lexicon Model for Ontologies [10] et fournit une base linguistique riche pour les ontologies, telles que la représentation des propriétés morphologiques et syntaxiques des entrées lexicales. Ce modèle s'inspire largement des modèles de données lexicaux précédents, en particulier LexInfo [11], LMF [12] et LIR [13], avec des améliorations telles qu'être purement en *Resource Description Framework* (RDF), ce qui le rend descriptif et modulaire et justifie sa promesse d'adaptabilité dans la gestion des ressources linguistiques. La figure 1 montre la conceptualisation de base de ce modèle qui est fondé sur le principe de référence sémantique où une entrée lexicale est définie par un individu, une classe ou une propriété définis dans l'ontologie.

3 Convertir le TLFi en Ontolex-Lemon

Le Trésor de la Langue Française est une des plus importantes ressources lexicographiques du français. Il contient 100 000 entrées, 270 000 définitions et 430 000 exemples du XIVème au XXème siècle [14]. La version informatisée de ce dictionnaire, appelée le TLF informatisé (TLFi) [2], est disponible sous format XML avec une DTD associée. La micro-structure de ce dictionnaire est enrichie par plusieurs types d'informations, notamment des sens et des définitions, des exemples d'usage, des étymologies, des indications d'emplois et de domaine général, ainsi que des locutions. En outre, les sens de chaque entrée peuvent être représentés dans une hiérarchie où les sens peuvent avoir des sous-sens pour montrer un sens plus strict. De ce fait, la structure de chaque entrée lexicale présente une complexité qui fait obstacle à son intégration dans des applications en TAL utilisant les standards actuels.

Pour faciliter l'utilisation du TLFi, nous l'avons donc converti au modèle Ontolex-Lemon. Afin de réaliser cette tâche, la structure XML du dictionnaire est parcourue et les éléments essentiels en sont extraits. Étant donné la complexité des données TLFi en XML et le manque d'uniformité dans la structure, l'extraction des données se concentre actuellement uniquement sur les lemmes, les catégories grammaticales, les sens et les définitions. Ces informations sont ensuite converties en format RDF en Ontolex-Lemon. Actuellement, 32 073 entrées du TLFi sont converties. L'annexe 2 présente le résultat de cette conversion pour l'entrée « baroque » (adjectif).

4 Conclusion et travaux futurs

Les standards actuels de données liées permettent d'augmenter l'inter-opérabilité et l'accessibilité des données langagières. Par conséquent, une version du TLFi en Ontolex-Lemon pourrait en permettre une meilleure intégration au sein des applications de TAL. La ressource est cependant très riche et nous travaillons à en extraire davantage de données pour pouvoir la convertir entièrement en Ontolex-Lemon dans un futur proche.

2. https://www.atilf.fr/ressources/tlfi/

Références

[1] Eric Laporte. Dictionaries for language processing. Readability and organization of information. PPGEL/UFES, 2013.

[2] Rada Mihalcea. Knowledge-based methods for WSD. In *Word sense disambiguation*, pages 107–131. Springer, 2007.

[3] Jie Zhou and Wei Xu. End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, pages 1127–1137, 2015.

[4] Matthieu Constant and Joakim Nivre. A transition-based system for joint lexical and syntactic analysis. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 161–171, 2016.

[5] Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia. *Linguistic Linked Data - Representation, Generation and Applications*, pages 3–9. Springer International Publishing, Cham, 2020.

[6] Rinke Hoekstra, Albert Meroño-Peñuela, Kathrin Dentler, Auke Rijpma, Richard Zijdeman, and Ivo Zandhuis. An ecosystem for linked humanities data. In *European Semantic Web Conference*, pages 425–440. Springer, 2016.

[7] John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. The Ontolex-Lemon model : development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21, 2017.

[8] Andon Tchechmedjiev. *Interopérabilité sémantique multilingue des ressources lexicales en données lexicales liées ouvertes*. PhD thesis, Université Grenoble Alpes, 2016.

[9] Sabine Tittel and Christian Chiarcos. Historical lexicography of old french and linked open data : Transforming the resources of the dictionnaire étymologique de l'ancien français with ontolex-lemon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). GLOBALEX Workshop (GLOBALEX-2018), Miyazaki, Japan*, pages 58–66, 2018.

[10] John McCrae, Dennis Spohr, and Philipp Cimiano. Linking lexical resources and ontologies on the semantic web with lemon. In *Extended Semantic Web Conference*, pages 245–259. Springer, 2011.

[11] Philipp Cimiano, Paul Buitelaar, John McCrae, and Michael Sintek. Lexinfo : A declarative model for the lexicon-ontology interface. *Web Semantics : Science, Services and Agents on the World Wide Web*, 9(1) :29–51, 2011.

[12] Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, and Claudia Soria. Lexical markup framework (lmf). In *International Conference on Language Resources and Evaluation-LREC 2006*, page 5, 2006.

[13] Elena Montiel-Ponsoda, Guadalupe Aguado De Cea, Asunción Gómez-Pérez, and Wim Peters. Modelling multilinguality in ontologies. *Coling 2008 : Companion volume : Posters*, pages 67–70, 2008.

[14] Ruth Radermacher. *Le Trésor de la langue française : une analyse lexicographique*. PhD thesis, Strasbourg 2, 2004.

A Le TLFi en Ontolex-Lemon (extrait)

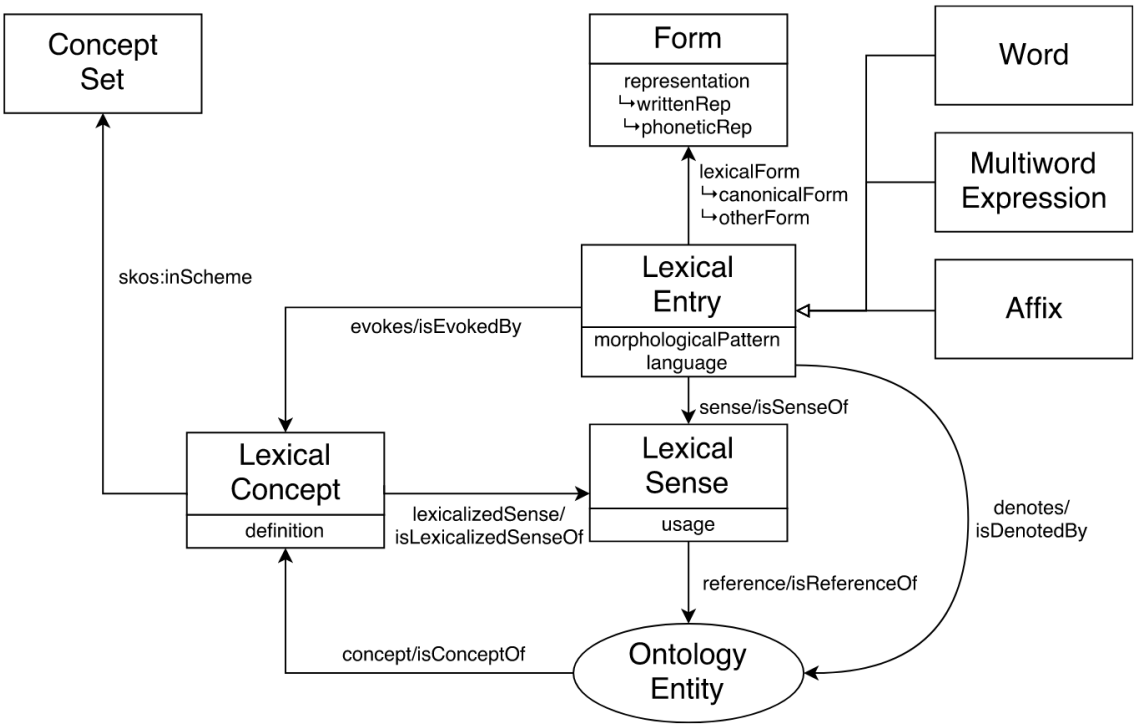


FIGURE 1 – Le modèle de données de base d'Ontolex-Lemon [7]

```
<https://www.cnrtl.fr/definition/11597> a ontolex:LexicalEntry ;
   rdfs:label "baroque"@fr ;
   lexinfo:partOfSpeech lexinfo:adjective ;
   ontolex:sense <https://www.cnrtl.fr/definition/11597#A._1.UND-1>,
       <https://www.cnrtl.fr/definition/11597#A._1.UND-2>,
       <https://www.cnrtl.fr/definition/11597#A._1.UND-3>,
       <https://www.cnrtl.fr/definition/11597#A._3.UND-6>,
       <https://www.cnrtl.fr/definition/11597#UND-9>.
   <https://www.cnrtl.fr/definition/11597#A._1.UND-1> skos:definition
       "Qui est caractéristique de la période qui a suivi la
       Renaissance classique."@fr .
   <https://www.cnrtl.fr/definition/11597#A._1.UND-2> skos:definition
       "Qui est caractéristique de la période musicale propre à
       l'Allemagne, àl'Angleterre, àl'Italie, et qui s'étend de 1580 à
       1760."@fr .
   <https://www.cnrtl.fr/definition/11597#A._1.UND-3> skos:definition
       "Qui appartient àl'époque littéraire qui, en France, correspond
       aux règnes de Henri IV et Louis XIII."@fr .
   <https://www.cnrtl.fr/definition/11597#A._3.UND-6> skos:definition
       "Artiste dont le style rappelle cette période"@fr .
   <https://www.cnrtl.fr/definition/11597#UND-9> skos:definition "Qui
       est de forme irrégulière, d'une rondeur imparfaite"@fr .
```

FIGURE 2 – La conversion de l'entrée « baroque » (adjectif) du TLFi en Ontolex-Lemon. L'entrée originale est accessible à `https://www.cnrtl.fr/definition/baroque`

Deux corpus audio transcrits de langues rares (japhug et na) normalisés en vue d'expériences en traitement du signal

Benjamin Galliot[2] Guillaume Wisniewski[3] Séverine Guillaume[2]
Laurent Besacier[4] Guillaume Jacques[5] Alexis Michaud[2] Solange Rossato[1]
Minh-Châu Nguyên[1, 2] Maxime Fily[2]

(1) Laboratoire d'Informatique de Grenoble (LIG), Unité Mixte de Recherche 5217 CNRS - Université Grenoble Alpes - Grenoble INP - Institut national de recherche en informatique et en automatique (INRIA)
(2) Langues et Civilisations à Tradition Orale (LACITO), Unité Mixte de Recherche 7107 CNRS - Sorbonne Nouvelle - Institut National des Langues et Civilisations Orientales (INALCO)
(3) Laboratoire de Linguistique Formelle (LLF), Unité Mixte de Recherche 7110 CNRS - Université de Paris
(4) Naver Labs Europe, Grenoble
(5) Centre de Recherches Linguistiques sur l'Asie Orientale (CRLAO), Unité Mixte de Recherche 8563 CNRS - École des Hautes Études en Sciences Sociales - Institut National des Langues et Civilisations Orientales

b.g01lyon@gmail.com, Guillaume.Wisniewski@univ-paris-diderot.fr, severine.guillaume@cnrs.fr, laurent.besacier@univ-grenoble-alpes.fr, rgyalrongskad@gmail.com, alexis.michaud@cnrs.fr, Solange.Rossato@univ-grenoble-alpes.fr, minhchau.ntm@gmail.com, maxime.fily@gmail.com

RÉSUMÉ

Deux corpus audio transcrits de langues « rares » (langues minoritaires de Chine : japhug et na) sont proposés comme corpus de référence pour des expériences en traitement automatique des langues. Les données, collectées et transcrites au fil d'enquêtes de terrain en immersion, s'élèvent à un total de 1907 minutes d'audio transcrit en japhug et de 209 minutes en na. Nous décrivons les traitements effectués pour les mettre à disposition sous une forme aisément accessible et utilisable, et présentons un outil qui permet d'assembler divers jeux de données de la collection Pangloss (archive ouverte de langues rares) en assurant la reproductibilité des expériences menées sur ces données.

ABSTRACT

Two Very-Low-Resource Language Speech Corpora for Experiments in NLP : Japhug and Na

Two audio corpora of minority languages of China (Japhug and Na), with transcriptions, are proposed as reference data sets for experiments in Natural Language Processing. The data, collected and transcribed in the course of immersion fieldwork, amount to a total of 1,907 minutes in Japhug and 209 minutes in Na. By making them available in an easily accessible and usable form, we hope to facilitate the development and deployment of state-of-the-art NLP tools for the full range of human languages. We present a tool for assembling datasets from the Pangloss Collection (an open archive) in a way that ensures full reproducibility of experiments conducted on these data.

MOTS-CLÉS : Corpus de référence, documentation computationnelle des langues, langues rares.

KEYWORDS: Benchmark datasets, Computational Language Documentation, low-resource languages, endangered languages.

1 Introduction

Le déploiement d'outils de traitement automatique de la parole comporte des enjeux évidents pour la documentation des langues, à une époque où le déclin de la diversité linguistique s'accélère (parallèlement au déclin de la biodiversité). Inversement, les langues rares présentent à la recherche en informatique tout un éventail de défis dont l'intérêt est de plus en plus clairement perçu.

Dans ce contexte, la mise à disposition de corpus de langues rares aisément accessibles, clairement versionnés et faciles d'utilisation paraît une nécessité tout à fait centrale. Dans le droit fil de la publication du corpus mbochi (bantou), décrite par Godard et al. (2018), nous avons déposé dans Zenodo deux corpus audio (avec transcriptions) de langues rares : le japhug et le na, langues minoritaires de Chine, de la famille sino-tibétaine.

Ces corpus ont été utilisés dans des travaux innovants en reconnaissance automatique de la parole (Adams et al., 2018, 2021; Macaire, 2021) et dans des réflexions interdisciplinaires associant talistes et linguistes (Michaud et al., 2018, 2019, 2020). Les corpus sont disponibles en ligne dans la collection Pangloss [1] (Michaud et al., 2016), une archive ouverte de langues rares hébergée par la plateforme Cocoon [2], mais les transcriptions de ces corpus sont enrichies et revues au fil des années, et de nouveaux documents s'y ajoutent, de sorte que renvoyer simplement au *corpus de langue L dans Pangloss* ne constitue pas une référence suffisamment précise pour parvenir à une reproductibilité d'expériences de TAL (ou d'autre type) menées sur ces données. Il s'agit, sinon de « données chaudes », du moins de « données tièdes », qui évoluent lentement au fil du temps.

Nous avons donc effectué un dépôt dans Zenodo d'un état donné de ces deux corpus, ainsi rendu accessible en quelques clics, sous une forme stabilisée (§2). Pour les collègues qui souhaiteraient aller au-delà d'une utilisation en l'état de ces deux jeux de données proposés au statut de *corpus de référence*, un outil est proposé (§3) pour panacher à son aise parmi l'ensemble des collections tout en conservant une garantie de reproductibilité, grâce au système de versionnage dont bénéficient les documents de la plateforme Cocoon.

2 Les dépôts : liens d'accès et choix techniques

Lieu de dépôt Les deux corpus ont été téléversés sur Zenodo, dont ils constituent respectivement les dépôts 5336698 (na) et 5521112 (japhug). Un corpus entier est identifié par un DOI (*digital object identifier*) : `10.5281/zenodo.5336698` pour le na, et `10.5281/zenodo.5521112` pour le japhug.

Le même type d'identifiant a été déployé pour la collection Pangloss, l'archive ouverte où sont déposés les corpus, mais avec une granularité tout à fait différente : un DOI pour chaque document (Vasile et al., 2020), ce qui est bien adapté pour les linguistes qui souhaitent faire référence aux données avec une granularité fine (un texte et, à l'intérieur d'un texte, un énoncé précis) mais ne donne pas prise sur un corpus entier.

Fichiers audio Les fichiers audio ont été dégradés en 16 bit, 16 kHz, mono. Là aussi, la logique qui préside à la constitution de ces corpus versionnés pour expériences de TAL s'éloigne de celle

1. `https://pangloss.cnrs.fr/`
2. `https://cocoon.huma-num.fr/`

de l'archivage pérenne dans la collection Pangloss. La taille des deux jeux de données déposés dans Zenodo est compatible (au jour d'aujourd'hui) avec des expériences menées sur un ordinateur portable : 1,8 Go pour le na, 9,2 Go pour le japhug.

Annotations Les annotations sont dans le format d'origine : du XML organisé selon une hiérarchie simple (un texte est composé de phrases, composées de mots, composés de morphèmes). Un prétraitement élémentaire a été effectué, afin de ne pas imposer aux utilisateurs de devoir prendre connaissance d'un certain nombre de conventions choisies par les déposants. En particulier, lors de la transcription de textes, il arrive que des retouches soient apportées, qui éloignent la transcription de ce qui a été dit sur l'enregistrement ; les passages ajoutés sont signalés par des crochets `[]`, et les passages que les consultants linguistiques souhaitent voir retranchés de la transcription « lissée » sont placés entre chevrons <>. Lors du prétraitement, les premiers ont été effacés, et les seconds allégés de leurs chevrons, afin qu'audio et transcription coïncident.

3 Un outil pour constituer de nouveaux jeux de données

Au-delà des deux jeux de données déposés dans Zenodo, il est bien sûr possible de constituer toutes sortes de nouveaux jeux de données à partir de la collection Pangloss. L'outil élaboré à l'occasion de la préparation des deux corpus déposés dans Zenodo, sobrement intitulé OutilsPangloss [3], consiste en une boîte à outils divers (en langage Julia) servant notamment à créer des (sous-)corpus de langues rares de Pangloss.

L'utilisateur·rice remplit un fichier YAML (dont différents exemples sont fournis), en indiquant notamment le nom de la langue. Elle peut également fournir une liste d'expressions rationnelles de modifications si des traitements sur les annotations sont à faire (telles que suppressions ou réarrangements de blocs de textes). Il est possible de filtrer par locuteur pour les sous-corpus. Des traitements sur l'audio peuvent également être paramétrés, pour choisir le taux d'échantillonnage et la profondeur, et séparer les différentes pistes des fichiers multicanaux en fichiers mono (démultiplexage).

Après le moissonnage (en Sparql), les vérifications des données par les métadonnées (hachage, versions, etc.) et les téléchargements, un fichier récapitulatif général (`données.yml`) se trouvera dans le dossier cible, aux côtés de dossiers `données` et `métadonnées`. Les informations contenues dans ce fichier récapitulatif suffisent à reproduire exactement, à tout moment, une expérience menée avec le jeu de données qu'il décrit, de sorte qu'il n'est pas techniquement nécessaire de travailler avec le jeu de données lui-même dans Zenodo.

Paradoxalement, si l'objet premier de la présente communication est d'attirer l'attention vers des jeux de données mis à disposition dans Zenodo, notre espoir serait que les pratiques s'orientent à l'avenir vers une description des jeux de données via des métadonnées renvoyant vers un unique hébergement des données dans une archive pérenne. En effet, décrire de cette façon le jeu de données qu'on a utilisé ne prend que quelques kilooctets (ko), tandis qu'un dépôt *en dur* de chaque bouquet de données multiplierait des dépôts (dans Zenodo ou ailleurs) dont chacun se compte en gigaoctets (Go).

On se contentera, en guise de mot de la fin, de relever que c'est aux différents acteurs de nos domaines de recherche qu'il appartient de s'emparer de ces outils, et de contribuer à façonner, par leurs choix, les directions que prendront les pratiques à l'interface du TAL et de la documentation des langues rares, dans le contexte d'une transition en cours vers la science ouverte.

3. `https://gitlab.com/lacito/outilspangloss`

Remerciements

Un grand merci aux collègues et amis consultants de langue japhug (en particulier Tshendzin) et na (en particulier M^me^ Latami Dashilame et son fils Latami Dashi). Le présent travail est une contribution au projet « La documentation computationnelle des langues à l'horizon 2025 » (ANR-19-CE38-0015-04) ainsi qu'au Labex « Fondements empiriques de la linguistique » (ANR-10-LABX-0083). Nous remercions l'Institut des langues rares (ILARA) de l'École pratique des hautes études, l'Université du Queensland et l'*Australian Research Council Centre of Excellence for the Dynamics of Language* pour le soutien financier apporté au développement d'outils logiciels pour la documentation linguistique.

Références

Adams, O., Cohn, T., Neubig, G., Cruz, H., Bird, S., and Michaud, A. (2018). Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *Proceedings of LREC 2018 (Language Resources and Evaluation Conference)*, pages 3356–3365, Miyazaki.

Adams, O., Galliot, B., Wisniewski, G., Lambourne, N., Foley, B., Sanders-Dwyer, R., Wiles, J., Michaud, A., Guillaume, S., Besacier, L., Cox, C., Aplonova, K., Jacques, G., and Hill, N. (2021). User-friendly automatic transcription of low-resource languages : plugging ESPnet into Elpis. In *Proceedings of ComputEL-4 : Fourth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, Hawai'i.

Godard, P., Adda, G., Adda-Decker, M., Benjumea, J., Besacier, L., Cooper-Leavitt, J., Kouarata, G. N., Lamel, L., Maynard, H., and Mueller, M. (2018). A very low resource language speech corpus for computational language documentation experiments. In *Proceedings of LREC 2018 (Language Resources and Evaluation Conference)*, pages 3366–3370, Miyazaki.

Macaire, C. (2021). Recognizing lexical units in low-resource language contexts with supervised and unsupervised neural networks. Master's thesis, Université de Lorraine.

Michaud, A., Adams, O., Cohn, T., Neubig, G., and Guillaume, S. (2018). Integrating automatic transcription into the language documentation workflow : experiments with Na data and the Persephone toolkit. *Language Documentation and Conservation*, 12 :393–429. Dans HAL : https ://halshs.archives-ouvertes.fr/halshs-01841979/.

Michaud, A., Adams, O., Cox, C., and Guillaume, S. (2019). Phonetic lessons from automatic phonemic transcription : preliminary reflections on Na (Sino-Tibetan) and Tsuut'ina (Dene) data. In *Proceedings of ICPhS XIX (19th International Congress of Phonetic Sciences)*, Melbourne.

Michaud, A., Adams, O., Cox, C., Guillaume, S., Wisniewski, G., and Galliot, B. (2020). La transcription du linguiste au miroir de l'intelligence artificielle : réflexions à partir de la transcription phonémique automatique. *Bulletin de la Société de Linguistique de Paris*, 116(1).

Michaud, A., Guillaume, S., Jacques, G., Mac, D.-K., Jacobson, M., Pham, T.-H., and Deo, M. (2016). Contribuer au progrès solidaire des recherches et de la documentation : la Collection Pangloss et la Collection AuCo. In *Journées d'Etude de la Parole 2016*, volume 1, pages 155–163.

Vasile, A., Guillaume, S., Aouini, M., and Michaud, A. (2020). Le Digital Object Identifier, une impérieuse nécessité ? L'exemple de l'attribution de DOI à la Collection Pangloss, archive ouverte de langues en danger. *I2D - Information, données & documents*, 2 :156–175.

DinG – a corpus of transcriptions of real-life, oral, spontaneous multi-party dialogues between French-speaking players of *Catan*

Maria Boritchev* Maxime Amblard*
(*) LORIA, UMR 7503, Université de Lorraine, CNRS, Inria, 54000 Nancy, France
{maxime.amblard,maria.boritchev}@loria.fr

MOTS-CLÉS : corpus, dialogue, transcription, questions, oral, français.

KEYWORDS: corpus construction, dialogue, transcription, multilogue, questions, oral, French.

1 General presentation of the corpus

We introduce a new corpus of manual transcriptions of real-life, oral, spontaneous multi-party dialogues between French-speaking players of *Catan*[1], called Dialogues in Games (DinG), first presented in (Boritchev, 2021). *Catan* is a board game for three to four players in which the main goal for each participant is to make their settlement prosper and grow, using resources that are scarce. Bargaining over these resources is a major part of the gameplay and constitutes the core of DinG's data. The corpus has been designed to showcase the SLAM corpus (Amblard et al., 2014a,b, 2015), a corpus of interviews of patients with schizophrenia, while being widely available.

Dialogues from DinG are unconstrained, as the players don't have to follow any rule or specific guideline, apart from playing the game. As bargaining over the resources is part of the gameplay, the players have to speak in order to play, so the dialogues are the ones naturally occurring in this particular setting. As the players have to speak to play, they do not discuss personal subjects outside the game setting, which makes it possible to completely anonymize the corpus by removing the players' names (de-identification).

The recordings took place during university game nights. As we wanted the participants to feel as relaxed and natural as possible, the recordings were conducted in the room where the rest of the game night took place. Recording during the game nights raised some technical challenges, in particular, because different people were playing different games in the same physical space. Yet, it allowed us to record in a way that made the participants very comfortable: most of them report afterward that they forgot the recording devices after the first fifteen minutes of playing. All recordings were conducted by a non-player observer, whose duties were to explain the experiment, find volunteers and supervise the smooth running of the process. In particular, the observer had to manage the microphone and monitor the level of surrounding noise. We needed to address the technical challenge of recording our participants in a clear enough way for transcription, without recording other people's conversations. In order to do so, we used *H2 next handy recorder* by ZOOM[2], in XY (90° recording mode) setting.

Some of the participants knew each other as friends and/or colleagues, but in most of the games

[1]Copyright ©2017 CATAN Studio, Inc. and CATAN GmbH. All rights reserved.

[2]https://zoomcorp.com/en/us/handheld-recorders/handheld-recorders/h2n-handy-recorder/

at least one player did not know the others at all. All participants are native French speakers. 33 people participated in the recording process, 12 women and 21 men. All participants but 3 had a master's degree or higher. Each participant only appears once in the corpus. We collected as little personal data as possible, but we can say that the average age of the participants is around 25 years old, and all the participants are native French speakers. All the participants signed an informed consent sheet, acknowledging they were giving us the right to record personal data (their voices) and share transcriptions of it.

The corpus was transcribed by paid transcribers, resulting in a high quality transcription. 6 transcribers took part in the project. 5 of them were recruited among natural language processing students, one is an expert in production and synchronization of subtitles. The transcription guide sets the norms to follow. The guide is an adaptation of (Blanche-Benveniste and Jeanjean, 1987). The main modifications are adaptations to the subject of our observation and the object of our research: (1) we specified the noise tags in order to adapt them to the board game context by adding tags such as [dice], [tokens]; (2) we added an explicit transcription of interrogative marks in order to account for utterances that were perceived (by the transcribers) as questions (rising intonation, answers given in the following dialogue turns). The transcribers who participated in the project have all received training on the same 5 minutes excerpt. Everyone did an individual segmentation and transcription before pooling and comparing the results.

The inter-annotator agreement for transcriptions is calculated on the transcription of a 5 minutes excerpt of DinG2, pre-segmented. Two independent annotators[3] (not working on the project before) have received empty segments for the excerpt and filled them with transcriptions, following the transcription guide. First, we computed the agreements for the full transcriptions, see the first two lines of table 1. It is important to stress that inter-annotator agreement on transcriptions is always low, as the amount of possible transcriptions is very large; yet, even taking this into account, the results we got were very low (under 0.3). Then, we computed the agreements for the transcriptions from which we removed the noises and the pauses. This produced lines 3 and 4 of table 1, with results higher than 0.5, which is usually considered to be a good agreement for transcriptions. This difference leads us to the conclusion that the quality of the recordings might be insufficient to grant an objective transcription of noises, on one hand, and also that transcriptions of the duration of pauses can vary from one transcriber to another.

	κ_{ipf}	**Raw agreement**
With noise, including unlinked/unmatched annotations	0.28	0.28
With noise, excluding unlinked/unmatched annotations	0.28	0.28
Without noise, including unlinked/unmatched annotations	0.52	0.55
Without noise, excluding unlinked/unmatched annotations	0.53	0.55

Table 1: Interrater agreement for transcription before/after noise tags and pauses removal, calculated with ELAN, following (Holle and Rein, 2013).

Transcribers are asked to respect scrupulously what is recorded/heard/said – they are not supposed to correct the language but to produce a faithful written version of French as it is used by the speakers. The writing of onomatopeias is normalized via lists (« euh », « hum », *etc.*). Pauses are explicitly marked with their approximate duration (ex: `(0.2s)`). Dysfluencies are also kept, in particular repetitions and beginning of words, that are marked with a hyphen (ex: « ca-(interrup-) carte » //

[3]We thank greatly Amandine Lecomte and Samuel Buchel for their contribution to our work.

"ca-(interrup-) card"). The transcription does not contain any punctuation except the interrogation point, which is used to annotate rising intonations that correspond to questions in the recording's oral context. As we are interested in questions and answers in dialogue, having an explicit annotation of questions is particularly useful for us.

It was of major importance for us to be able to distribute our resource while preserving the participants' private data. The last step in the transcription process is anonymizing the transcription. Each of the players is identified with the colour of their game pieces: Red (**R**), White (**W**), Yellow (**Y**) or Blue (**B**). If a name is pronounced out loud, it is replaced in the transcription by the name of the corresponding color, in upper case. Outside noises and speakers are assigned to an outside speaker called Other (**O**).

An average game of *Catan* lasts at least 30 minutes, thus DinG contains long interactions, going beyond informative exchanges. The corpus was originally designed to study human-human dialogue based on attested, spontaneous, and unconstrained oral data in French. Its nature allows for large dissemination and high cross-domain reusability. Its length allows for a study from different perspectives. The following shows an excerpt from the corpus[4]:

```
009 Y   j'aimerais bien faire 7 pour une fois
         00:00:14.438 - 00:00:15.880
         (0.64)
```

009 Y I would like to get a 7 for once

```
010 R   en fait t'as (te-) t'étais contente parce que juste tu as
fait un double 6 et qu'en général c'est cool dans les jeux [rire]
         00:00:16.518 - 00:00:21.910
```

010 R in fact your have (y-) you were happy because simply you got a double 6 and generally it's cool in games [laugh]

```
011 Y   ouais c'est ça
         00:00:21.712 - 00:00:22.718
```

011 Y yeah that's it

The corpus is available on Gitlab: `https://gitlab.inria.fr/semagramme/DinG`. It is distributed under the Attribution ShareAlike Creative Commons license (CC BY-SA 4.0). Each game is available as a numbered .txt file, exported from ELAN[5] (Wittenburg et al., 2006).

2 Corpus description

DinG is composed of 10 recordings of games that last 70 minutes on average. The shortest recording is almost 40 minutes long (DinG8), the longest lasts a little over 1h44m (DinG1). Table 2 shows the first corpus measurements.

DinG1 is the longest both with respect to time and amount of speech turns; it also contains the biggest amount of questions. While DinG9 and 10 are not the shortest in terms of time, their amount of

[4]The participants are designated by the colour of their tokens: **Red** (**R**), **White** (**W**), **Yellow** (**Y**), **Blue** (**B**).
[5]`https://archive.mpi.nl/tla/elan`

Name	Length (min)	Length (turns)	# questions	# turns /minute	# questions /minute	% questions among turns
DinG1	**104.33**	**3,572**	**506**	34.24	4.85	**14.17**
DinG2	86.31	2,969	290	34.40	3.36	9.77
DinG3	53.7	1,716	126	31.96	2.35	**7.34**
DinG4	75.93	2,985	333	39.31	4.39	11.16
DinG5	78.41	3,012	362	38.41	4.62	12.02
DinG6	84.02	3,130	265	37.25	3.15	8.47
DinG7	96.34	3,293	340	34.18	3.53	10.32
DinG8	**39.92**	1,627	196	**40.76**	**4.91**	12.05
DinG9	41.71	**795**	**69**	**19.06**	**1.65**	8.68
DinG10	41.13	**476**	**41**	**11.57**	**1.00**	8.61
Global data	701.8	23,575	2,528	33.59	3.60	10.72
CV	34%	47%	57%	29%	40%	20%

Table 2: DinG data – observations per game, on average and coefficients of variation (CV).

speech turns and questions are significantly (more than 10%) smaller than DinG8's (shortest in terms of time). This observation is supported by the fact that DinG9 and 10 present the smallest amount of speech turns per minute, while DinG8 presents the greatest: DinG8 lasts less time but DinG8's players talked at least twice more than DinG9 and DinG10's ones. Similarly, DinG8 presents the highest amount of questions per minute while DinG9 and DinG10 show the smallest ones.

The focus returns on DinG1 when we look at the percentage of questions among all the speech turns, as this game presents the highest percentage (the smallest one is shown by DinG3). DinG is homogeneous in terms of all the measures used in table 2, as all the coefficients of variation stay under 60%. While the amount of questions (the utterances marked with a '?') varies quite a lot from one recording to another, the percentage of questions among turns stays very similar (under 30%).

3 Future perspectives

We envision three perspectives for further development of DinG: its extension, its annotation, and its usage. A first step would be to transform it to fit the TEI format[6] (Parisse and Liégeois, 2020). Another path we envision is through the anonymization of the recordings through approaches such as the ones described in (Qian et al., 2017). Once the transcriptions, the oral data, and the participants' consent are available, a synchronization work would have to take place to enrich the resource. Then, transcription constitutes a first level of linguistic annotation. We would like to offer other annotations, at different linguistic levels: morphosyntactic, part-of-speech, disfluencies, syntactic (through universal dependencies, for example). We would also like to annotate on layers specific to dialogue: dialogue transactions, connectives, argumentation structures, in particular, throughout the annotation schemata that were developed for the STAC project (Asher et al., 2016).

Finally, this corpus can be used as a starting point for fine-grained analysis on the mechanisms underlying the articulations of questions and answers in French, such as the ones presented in (Boritchev and Amblard, 2021). A first step would be the inclusion of DinG in the French Question banks (Judge et al., 2006; Seddah and Candito, 2016).

[6] https://tei-c.org/Guidelines/

References

Amblard, M., Fort, K., Demily, C., Franck, N., and Musiol, M. (2015). Analyse lexicale outillée de la parole transcrite de patients schizophrènes. *Traitement Automatique des Langues*, 55(3):91 – 115.

Amblard, M., Fort, K., Musiol, M., and Rebuschi, M. (2014a). L'impossibilité de l'anonymat dans le cadre de l'analyse du discours. In *Journée ATALA éthique et TAL*, Paris, France.

Amblard, M., Musiol, M., and Rebuschi, M. (2014b). L'interaction conversationnelle à l'épreuve du handicap schizophrénique. *Recherches sur la philosophie et le langage*, 31:1–21.

Asher, N., Hunter, J., Morey, M., Benamara, F., and Afantenos, S. (2016). Discourse Structure and Dialogue Acts in Multiparty Dialogue: the STAC Corpus. In *10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2721–2727, Portoroz, Slovenia.

Blanche-Benveniste, C. and Jeanjean, C. (1987). *Le français parlé: transcription et édition*. Didier érudition.

Boritchev, M. (2021). *Dialogue Modeling in a Dynamic Framework*. PhD thesis.

Boritchev, M. and Amblard, M. (2021). Picturing questions and answers—a formal approach to slam. In *(In) coherence of Discourse*, pages 65–89. Springer.

Holle, H. and Rein, R. (2013). The modified Cohen's kappa: Calculating interrater agreement for segmentation and annotation. *Understanding Body Movement: A Guide to Empirical Research on Nonverbal Behaviour, H. Lausberg, Ed. Frankfurt am Main: Peter Lang Verlag*, pages 261–277.

Judge, J., Cahill, A., and Van Genabith, J. (2006). Questionbank: Creating a corpus of parse-annotated questions. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 497–504.

Parisse, C. and Liégeois, L. (2020). Utiliser les outils CORLI de conversion TEI pour l'analyse de corpus de langage oral. In *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 4: Démonstrations et résumés d'articles internationaux*, pages 64–65. ATALA; AFCP.

Qian, J., Du, H., Hou, J., Chen, L., Jung, T., Li, X.-Y., Wang, Y., and Deng, Y. (2017). Voicemask: Anonymize and sanitize voice input on mobile devices. *arXiv preprint arXiv:1711.11460*.

Seddah, D. and Candito, M. (2016). Hard time parsing questions: Building a questionbank for French. In *Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). Elan: a professional framework for multimodality research. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1556–1559.

Exploration de systèmes *end-to-end* pour la reconnaissance automatique de la parole spontanée

Solène Evain[1], Solange Rossato[1], Benjamin Lecouteux[1], François Portet[1]
(1) Univ. Grenoble Alpes, CNRS, Grenoble-INP, LIG, 38000 France
prenom.nom@univ-grenoble-alpes.fr

MOTS-CLÉS : Reconnaissance automatique de la parole, système end-to-end, parole spontanée.

KEYWORDS: Automatic speech recognition, end-to-end system, spontaneous speech.

Ces dernières années, les systèmes de Reconnaissance Automatique de la Parole (RAP) ont donné de très bons résultats sur les benchmarks de la communauté. Si ces résultats sont très bons sur la parole lue ou médiatique, les performances baissent considérablement pour la Reconnaissance de la Parole Spontanée (RAPS), notamment à cause de la faible disponibilité des corpus et de la difficulté de définir et de modéliser ce type de parole. Dans ce travail, nous souhaitons explorer l'utilisation d'un modèle neuronal pour la RAPS. En effet, l'optimisation *end-to-end* (de bout en bout) de ces modèles – sans modèle de langue *a priori* et en partie sans corpus annoté – offre non seulement des performances intéressantes, mais également l'opportunité d'étudier la modélisation de la parole spontanée uniquement à partir de données.

1 Vers une définition de la parole spontanée

Il est difficile de définir précisément ce qu'est la parole spontanée. D'une part, elle admet plusieurs dénominations : parole spontanée, *casual speech* (parole détendue) (Torreira et al., 2010), parole non scriptée (Llisterri, 1992), populaire (Guiraud, 1965), informelle, familière, non standard (Blanche-Benveniste, 1999), non préparée (Dufour et al., 2010), non planifiée (Veiga et al., 2012), non conventionnelle (Caron, 1992), conversation naturelle ou encore conversationnelle (Shriberg, 2005). D'autre part, Fujisaki (Fujisaki, 1997) parlait en 1997 d'un "continuum du degré de spontanéité", avec d'un côté la parole la plus préparée (la plus contrainte) et d'un autre la moins préparée (la plus libre). Derrière cette notion de continuum réside l'idée qu'il n'y a pas de limite franche entre parole préparée et parole spontanée. Enfin, certains articles font également état de niveaux de spontanéité, basés sur le nombre d'hésitations (Veiga et al., 2012) ou sur l'intelligibilité (Dufour et al., 2010). Le lien entre la situation d'énonciation, le degré d'intimité entre les locuteurs et le mode de communication (présentiel/distanciel, devant une foule etc) peut également permettre de déterminer un niveau de spontanéité. Ces trois points nous démontrent qu'il est difficile d'avoir une définition précise de ce qu'est la parole spontanée mais que l'on peut la lier au fait de parler sans contrainte (parole naturelle) et sans préparation (non préparée) et qui semble apparaître dans des contextes où peu d'attention est portée sur la forme du discours et où l'aspect conventionnel de la langue peut être omis. Luzzati (Luzzati, 2007) la définit comme le fait de "quitter l'univers de la phrase, celui de la langue préparée, celui de l'erreur qui s'efface et se rature, pour basculer du côté des énoncés non prémédités, dont l'émetteur est le premier auditeur, dans lesquels l'erreur se traduit par un allongement du message".

D'un point de vue caractéristique, la parole spontanée est notamment représentée par des disfluences (amorces de mots, hésitations, reprises, répétitions...) (Adda-Decker et al., 2004), une vitesse d'élocution assez rapide (Adda-Decker et al., 2012) ou à débit variable, une réduction temporelle fréquente (Wu and Adda-Decker, 2020), de nombreux allongements (Duez, 2001) et une prononciation peu soignée (hypoarticulation) (Dufour, 2008).

2 Les corpus de parole spontanée en français

Dans ce travail, nous avons recensé seize corpus du français comprenant de la parole spontanée, comme le montre le tableau 1. La durée totale de ces corpus est d'environ 1825 heures, ce qui est à

Corpus (date)	**Type parole**	**Type d'interaction**	**Durée** (approx.)
ESLO1 (1968-71)	Spont.	repas, entretiens	318 h
ESLO2 (2008-)	Spont.	repas, entretiens	450 h
NCCFr (2010)	Spont.		36 h
MPF (2010-14 ?)	Spont.	entretiens	78 h
PFC (1999)	Lue/Spont.	lecture de liste de mots/textes, entretiens, conversations libres	>300 h ?
CFPP2000 (2005- ?)	Spont.	entretiens	38 h* / 58 h 40
CLAPI (1998-)	Spont./Prep.	conversations, réunions, visites guidées	16 h 30*
C-ORAL-ROM (2001-2003)	Spont./Prep.		22 h*
CRFP (1998-2002)	Spont./Prep.	souvenirs, théâtre, émissions, cours	34 h*
FLEURON (2009-2012)	Prep./Spont.	Interactions étudiants/administration	3 h 25*
OFROM (2008-2012)	Spont./Prep.	discussions, entretiens, communications	25 h 13*
TCOF (2005-2009)	Spont./Prep.	discussions, entretiens, réunions, débats	28 h40* / 61 h
TUFS (2005-2011)	Spont.	interviews, entretiens, discussions	52 h 40*
CFPB (2013-2015)	Spont.	entretiens	5 h* / 21 h 30
Réunions (2007-2008)	Prep./Spont.	réunions	18 h*
Valibel (1998-2008)	Spont./Prep.	interviews, discours, entretiens, journaux, souvenirs, conversations...	43 h 25* / 331 h
Total			1825 h 58

*Compris dans le corpus CEFC ; 'Spont.' : spontanée, 'Prep.' : préparée

TABLE 1 – Corpus de parole spontanée en français

première vue tout à fait acceptable pour pouvoir apprendre des modèles de RAP. Cependant, cette durée globale n'est pas la durée effective de parole spontanée disponible. En effet, neuf corpus sur seize comprennent à la fois de la parole spontanée et de la parole préparée et/ou lue, sans que l'on sache quelle est la proportion de chacun de ces types de parole. De plus, l'accès à certains de ces corpus n'est pas toujours aisé. Hormis certaines contraintes administratives (signature d'un contrat), il existe des corpus pour lesquels les sites ne permettent pas le téléchargement groupé ce qui oblige à aller chercher les fichiers un à un. Certains sont difficilement accessibles pour des raisons d'anonymisation et de traitement/transcription en cours. Enfin, dans des cas extrêmes, le contact avec les chercheurs ayant constitué les corpus est difficile, voire impossible, dû à peu de disponibilité de leur part ou à des départs (certains corpus sont assez anciens) ce qui enlève tout accès aux données. Par ailleurs, il convient de noter que pour beaucoup de corpus, des outils et des conventions de

transcription différentes ont été employés, ce qui rend leur fusion fastidieuse. Enfin, d'un point de vue plus technique, le matériel ou les conditions d'enregistrement, font qu'une partie des enregistrements sont inexploitables pour un objectif de RAP.

3 RAP et importance des données

L'accès à des données audio annotées est primordial pour l'entraînement d'un système de RAP. L'étude de (Lamel et al., 2002) montre que le *Word Error Rate* ou *WER* (taux d'erreur de mots) est dépendant de la quantité de données d'apprentissage. Or, l'accès à une grande quantité de données n'est possible que pour quelques dizaines de langues sur les 7 000 existantes. Les bonnes performances observées aujourd'hui sont donc plus représentatives de la reconnaissance de langues telles que l'anglais ou le français. La reconnaissance de la parole lue, préparée et spontanée ne sont pas non plus équivalentes (Tancoigne et al., 2020). Là aussi, le manque de données est problématique, notamment pour la parole spontanée. Enfin, lorsque l'apprentissage et le décodage se font sur des données issues du même corpus, il est difficile d'évaluer réellement la capacité de généralisation du système. En effet, lors d'un décodage sur des données issues d'un corpus différent de celui d'apprentissage, le *WER* augmente (Likhomanenko et al., 2021).

4 Les limites d'un système à base de HMM pour la RAPS

Pour un système de RAP de type *HMM-GMM/DNN*, l'objectif est de trouver la séquence de mots la plus vraisemblable étant donnée une séquence de paramètres acoustiques. Un tel système, s'appuie sur un modèle acoustique, un lexique phonétisé et un modèle de langue donnant la probabilité d'une séquence de mots. Le poids accordé au modèle de langue dans ce type d'architecture est important. La parole spontanée étant différente de la parole lue ou préparée, sa reconnaissance nécessite d'adapter un système de RAP directement sur ce type de parole. À la fin des années 2000, la quantité de données nécessaires pour construire un modèle de langue était de plusieurs dizaines à plusieurs centaines de millions de mots et de plusieurs dizaines à plusieurs centaines d'heures (Pellegrini, 2008) pour l'adaptation d'un modèle acoustique. Or, les corpus de parole spontanée annnotés étant assez peu nombreux, les modèles de langue sont alors souvent construits sur d'autres types de données textuelles comme des journaux ou des transcriptions de parole lue ou préparée. La parole spontanée comprenant des disfluences (répétitions, allongements...) dues à la construction du message au cours de sa réflexion n'est alors pas représentée dans le modèle de langue ayant pourtant un fort impact sur le processus de décodage de la parole. En ce qui concerne le lexique phonétisé, la difficulté pour la RAPS réside dans la présence de nombreuses variantes lexicales (dues notamment au phénomène d'hypoarticulation) qu'il serait coûteux de toutes représenter.

5 L'apport des systèmes neuronaux

Étant donné le manque de données de parole spontanée transcrite en français et la complexité de traitement de ce type de parole, l'utilisation d'un système de RAP de type HMM-GMM/DNN ne semble pas le plus adéquat. Nous nous interrogeons donc aujourd'hui sur l'apport possible d'un

système *end-to-end* n'utilisant plus de lexique phonétisé et dont l'usage d'un modèle de langue externe devient optionnel : ceci permet d'injecter moins d'*a priori* sur la langue dans le système de RAP. De plus, l'association étant directe entre le signal en entrée et sa transcription en sortie (Chan et al., 2016) , le système apprend ses propres représentations. Néanmoins, ce type de système est très gourmand en données. Nous pensons étudier les systèmes pré-entraînés sur le français, tels que ceux appris pour LeBenchmark (Evain et al., 2021) pour pallier le manque de données d'apprentissage de parole spontanée. Cette technique consiste à apprendre des représentations globales de la langue de façon auto-supervisée, grâce à une grande quantité de données non-annotées. L'étude de (Baevski et al., 2020) montre que l'utilisation d'un modèle pré-entraîné suivi d'un ajustement du modèle sur 10 heures de parole annotées donne un meilleur WER (même sans modèle de langue) que l'utilisation d'un système HMM-DNN entraîné avec 100 heures de données annotées. Cette étude a été faite sur l'anglais, avec des modèles appris sur des corpus de lecture pour une tâche de RAP sur de la lecture. L'impact de l'utilisation de modèles pré-entraînés sur la parole spontanée en français reste donc à explorer et à analyser (Chung et al., 2020).

6 Perspectives

Ce travail a pour but d'explorer l'utilisation de systèmes de RAP *end-to-end* avec utilisation de modèles pré-entraînés sur le français pour la reconnaissance de la parole spontanée. Nous souhaitons évaluer l'influence du degré de spontanéité (ESLO2), du degré d'interaction (CRFP, Valibel, ESLO2) et du niveau d'intimité entre les locuteurs (ESLO2) sur la reconnaissance de cette parole. Pour cela, plusieurs corpus de test seront sélectionnés ou "composés" en respectant les critères suivants : premièrement, les données ne devront pas avoir été utilisées pour l'élaboration des modèles pré-entraînés. Ensuite, la qualité des fichiers audio doit être suffisante pour une tâche de RAP (nous ne nous focalisation pas sur la parole bruitée). Enfin, les enregistrements devront comprendre 3 locuteurs maximum, ceci afin d'éviter les situations de schismes interactionnels [1]. Par "composition" d'un corpus de test, nous entendons le rassemblement de fichiers audio de différents corpus. La grille d'analyse des résultats comprendra plusieurs niveaux : prosodique, morphologique et grammatical, afin d'étudier la gestion d'événements propres à la parole spontanée (impact de la segmentation des données en entrée, gestion des nouveaux mots et des suites de mots "non conventionnelles"). À des fins de comparaison, notre système sera également évalué sur un corpus de lecture (Commonvoice). En ce qui concerne le système de RAP, nous allons réutiliser l'architecture de type encodeur-décodeur utilisée dans la partie RAP de LeBenchmark (Evain et al., 2021) : *CRDNN (VGG-RNN-DNN) - Joint CTC/attention LSTM*, prenant en entrée des représentations de type Wav2vec.

7 Adéquation aux thématiques du GDR LIFT

La parole spontanée partage avec les langues peu dotées la caractéristique du manque de ressources. Nous espérons ainsi que les techniques d'analyse et d'apprentissage qui seront définies dans ce travail pourront ouvrir des perspectives nouvelles pour l'analyse linguistique, que ce soit pour collecter et annoter des données (systèmes de RAP) ou pour extraire ou vérifier des généralisations linguistiques (analyse des représentations apprises par le modèle).

1. Lorsqu'il y a plus de trois personnes en interaction, plusieurs conversations peuvent avoir lieu en parallèle.

Références

Adda-Decker, M., Fougeron, C., Gendrot, C., Delais-Roussarie, E., and Lamel, L. (2012). La liaison dans la parole spontanée familière : une étude sur grand corpus. *Revue francaise de linguistique appliquee*, Vol. XVII(1) :113–128. Bibliographie_available : 1 Cairndomain : www.cairn.info Cite Par_available : 1 Publisher : Publications linguistiques.

Adda-Decker, M., Habert, B., Barras, C., Adda, G., de Mareüil, P. B., and Paroubek, P. (2004). Une étude des disfluences pour la transcription automatique de la parole spontanée et l'amélioration des modèles de langage. In *Actes des 25èmes Journées d'Etudes sur la Parole (JEP 2004)*, Fès, Maroc.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0 : A Framework for Self-Supervised Learning of Speech Representations. In *Proceedings of the 34th conference on Neural Information Processing Systems (NeurIPS 2020)*, page 12, Vancouver, Canada.

Blanche-Benveniste, C. . a. (1999). "Français parlé - oral spontané". Quelques réflexions. *Revue française de linguistique appliquée*, IV(2) :21.

Caron, P. (1992). L'écriture de la noblesse vers 1680. In *Grammaire des fautes et français non conventionnels*. Paris, France, presses de l'ecole normale supérieure, rue d'ulm edition.

Chan, W., Jaitly, N., Le, Q., and Vinyals, O. (2016). Listen, attend and spell : A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964. ISSN : 2379-190X.

Chung, Y.-A., Tang, H., and Glass, J. (2020). Vector-Quantized Autoregressive Predictive Coding. In *Proceedings of Interspeech 2020*, pages 3760–3764, Shanghai, China. ISCA.

Duez, D. (2001). Signification des hésitations dans la productione t la perception de la parole spontanée. *Parole*, (17/18/19) :113–138.

Dufour, R. (2008). From prepared speech to spontaneous speech recognition system : a comparative study applied to French language. In *Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology*, CSTST '08, pages 595–599, New York, NY, USA. Association for Computing Machinery.

Dufour, R., Estève, Y., and Deléglise, P. (2010). Automatic indexing of speech segments with spontaneity levels on large audio database | Proceedings of the 2010 international workshop on Searching spontaneous conversational speech. In *SSCS '10 : Proceedings of the 2010 international workshop on Searching spontaneous conversational speech*, Firenze, Italy.

Evain, S., Nguyen, H., Le, H., Boito, M., Mdhaffar, S., Alisamir, S., Tong, Z., Tomashenko, N., Dinarelli, M., Parcollet, T., Allauzen, A., Estève, Y., Lecouteux, B., Portet, F., Rossato, S., Ringeval, F., Schwab, D., and Besacier, L. (2021). LeBenchmark : A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech. In *Proceedings of Interspeech 2021*, Brno, Czechia.

Fujisaki, H. (1997). Prosody, Models, and Spontaneous Speech. In *Computing Prosody*. Springer, New York, NY, sagisaka y., campbell n., higuchi n. (eds) edition.

Guiraud, P. (1965). *Le français populaire*. Number 1172 in "Que sais-je ?". Paris, France, presses universitaires de france edition.

Lamel, L., Gauvain, J.-L., and Adda, G. (2002). Unsupervised acoustic model training. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02)*, pages I–877–I–880.

Likhomanenko, T., Xu, Q., Pratap, V., Tomasello, P., Kahn, J., Avidov, G., Collobert, R., and Synnaeve, G. (2021). Rethinking Evaluation in ASR : Are Our Models Robust Enough ? In *Proceedings of Interspeech 2021*, pages 311–315, Brno, Czechia. ISCA.

Llisterri, J. (1992). Speaking styles in speech research. Dublin, Ireland.

Luzzati, D. (2007). Le dialogue oral spontané : quels objets pour quels corpora. *Revue d'Interaction Homme-Machine*, 8(2).

Pellegrini, T. (2008). *Transcription automatique de langues peu dotées*. PhD thesis, Université Paris Sud - Paris XI, Paris, France.

Shriberg, E. (2005). Spontaneous Speech : How People Really Talk and Why Engineers Should Care. In *Proceedings of Interspeech 2005*, Lisbon, Portugal.

Tancoigne, E., Corbellini, J.-P., Deletraz, G., Gayraud, L., Ollinger, S., and Valéro, D. (2020). La transcription automatique : un rêve enfin accessible ? Technical report, MATE-SHS.

Torreira, F., Adda-Decker, M., and Ernestus, M. (2010). The Nijmegen Corpus of Casual French. *Speech Communication*, 52(3) :201. Publisher : Elsevier : North-Holland.

Veiga, A., Candeias, S., Celorico, D., Proença, J., and Perdigão, F. (2012). Towards Automatic Classification of Speech Styles. In *Proceedings of the 10th international conference on Computational Processing of the Portuguese Language*, Coimbra, Portugal. Pages : 426.

Wu, Y. and Adda-Decker, M. (2020). Réduction temporelle en français spontané : où se cache-t-elle ? Une étude des segments, des mots et séquences de mots fréquemment réduits. In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)*, volume Volume 1 : Journées d'Études sur la Parole, pages 627–635, Nancy, France.

Le projet ANR Autogramm et l'extraction automatique de grammaires Illustration par la négation

Sylvain Kahane[1] Bruno Guillaume[2] Kim Gerdes[3] Bernard Caron[4] Sylvain Loiseau[5]
(1) Modyco, Université Paris Nanterre & CNRS
(2) Loria, Nancy
(3) LISN, Université Paris Saclay & CNRS
(4) Llacan, CNRS
(5) Lacito, Université Paris Nord & CNRS
sylvain@kahane.fr, bruno.guillaume@loria.fr, kim@gerdes.fr
bernard.caron@cnrs.fr, sylvain.loiseau@univ-paris13.fr

MOTS-CLES : Treebank, texte avec glose alignée, induction de grammaire, typologie quantitative

KEYWORDS: Treebank, Interlinear Glossed Text, grammar induction, quantitative typology

Nous nous proposons de présenter les objectifs du projet ANR Autogramm qui démarrera au 1er janvier 2022 (2022-2025). L'objectif du projet de d'extraire automatiquement des observations pertinentes à partir de corpus annotés, qu'il s'agisse de textes avec une glose alignée (Interlinear Glossed Texts) (ou de treebanks syntaxiques du type (Surface-syntactic) Universal Dependencies (UD et SUD) (de Marneffe et al. 2021, Gerdes et al. 2018)

Le projet prévoit une collaboration entre linguistes de terrain pour (une quinzaine de langues, la plupart sous-dotées, voire en danger), linguistes formels et talistes. Le projet sera l'occasion de stabiliser et développer la chaine de traitement de treebanks reposant actuellement sur ArboratorGrew pour l'annotation, Grew-match pour les requêtes et Grew pour la conversion et l'enrichissement automatique des treebanks (Guillaume 2021, Guibon et al. 2020).

Un autre volet du projet est d'extraire automatiquement des descriptions grammaticales des ressources et en particulier des observations quantitatives du genre « 98% des sujets en français sont avant le verbe », « 29% des sujets après le verbe se trouvent dans une relative », « 23% des sujets dans une relative sont après le verbe » et de classer ces observations par ordre de pertinence. La plupart des observations que nous voulons faire sont de la forme « le motif M1 implique le motif

M2 dans x% des cas (M1 => M2 @x%). Par exemple, « 98% des sujets en français sont avant le verbe » se traduit par (X-[subj]->Y) => (Y<<X) @98% (nous utilisons ici les conventions de Grew). A partir de deux motifs M1 et M2 tels que M1 => M2 @98% (et donc M1 => ¬M2 @2%), nous souhaitons induire automatiquement les facteurs M3 qui facilitent les exceptions (les 2% de ¬M2), c'est-à-dire extraire les motifs M3 pour lesquelles nous avons significativement plus de chances d'avoir M1 & ¬M2. Le fait que X soit la tête d'une relative (Z-[mod:relcl]->X) est le motif M3 le plus pertinent pour le cas du sujet inversé.

Un dernier volet du projet est de constituer une base typologique d'observations pour l'ensemble des langues pour lesquelles des ressources seront disponible (la base UD comporte actuellement des treebanks de plus de 140 langues). Il s'agit d'observations quantifiées qui induisent donc une méthodologie d'exploitation nouvelle (Gerdes et al. 2021).

Dans le cadre de l'appel du GDR LIFT, nous nous proposons de mettre à l'épreuve les objectifs du projet Autogramm par une étude de la négation dans différentes langues que nous avons déjà étudiées et pour lesquelles des treebanks sont disponibles et notamment le français avec sa double négation, le wolof et sa négation par un morphème flexionnel ou un auxiliaire (Bondéelle & Kahane 2021), le naija et sa négation par la particule *no* dont on peut montrer qu'elle se comporte comme un auxiliaire (Caron et al. 2019), etc.

Références

Bondéelle, O., & Kahane, S. (2021). Les particules verbales du wolof et leur combinatoire syntaxique et topologique. *Bulletin de la Société de Linguistique de Paris*, *115*(1), 391-465.

Caron, B., Courtin, M., Gerdes, K., & Kahane, S. (2019). A surface-syntactic UD treebank for Naija. In *Proceedings of Treebanks and Linguistic Theories,* Syntaxfest.

Comrie, Bernard, Haspelmath, Martin, & Bickel, Balthasar. 2008. The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses. Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology & the Department of Linguistics of the University of Leipzig. Retrieved January, 28, 2010. https://www.eva.mpg.de/lingua/resources/glossing-rules.php.

de Marneffe, M. C., Manning, C. D., Nivre, J., & Zeman, D. 2021. Universal dependencies. *Computational Linguistics*, 47(2), 255-308.

Gerdes, K., Guillaume, B., Kahane, S., & Perrier, G. 2018. SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of the Second Universal Dependencies Workshop (UDW)*.

Gerdes, K., Kahane, S., & Chen, X. (2021). Typometrics: From Implicational to Quantitative Universals in Word Order Typology. *Glossa: a journal of general linguistics*, *6*(1).

Guibon, G., Courtin, M., Gerdes, K., & Guillaume, B. (2020). When Collaborative Treebank Curation Meets Graph Grammars. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC).*

Guillaume, B. 2021. Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL)*, 168–175.

Le(s)? chinois du Shun-pao 申報

Pierre Magistry
ertim, INALCO, Paris
pierre.magistry@inalco.fr

MOTS-CLÉS : Corpus historique, AFC, plongements contextualisés, chinois mandarin, chinois classique.

KEYWORDS: Historical corpora, FCA, contextualized word embeddings, Mandarin Chinese, Classical Chinese.

1 Introduction

Les travaux présentés ici poursuivent un objectif double. Premièrement, il s'agit d'approfondir l'étude inédite que constitue l'exploration du premier journal quotidien imprimé en sinogrammes, le Shun-pao, dont le texte intégral numérisé n'était pas disponible pour des traitements massifs jusqu'à tout récemment. Ce corpus est un témoignage d'une pratique du chinois écrit au tournant du 20ème siècle. La période que couvre ce quotidien présente un intérêt tout particulier. C'est à cette époque que se construit et s'impose le projet d'une langue nationale unique pour la Chine, et que les formes vernaculaires prennent le dessus sur l'usage du chinois classique dans les pratiques écrites. L'accès au corpus de la totalité des textes numérisé rend possible des études quantitatives pour caractériser la diversité linguistique du corpus et son évolution sur une période allant de 1872 à 1949 en observant directement les usages.

Le second objectif de notre travail est méthodologique, en proposant une comparaison de plusieurs approches. On souhaite mettre en regard les apports de méthodes neuronales basées sur des plongements lexicaux contextualisés avec une approche plus classique fondée sur l'analyse factorielle des correspondances. On insistera sur les critères pris en compte pour favoriser une méthode plutôt qu'une autre, et sur les types de questionnements que l'on peut aborder avec l'une et l'autre des méthodes.

2 Les corpus

Cette étude se focalise sur le corpus du Shun-pao, mais nous avons aussi recourt à d'autres corpus comme points de comparaison. Il s'agit d'une part d'un corpus arboré de chinois moderne standard de Taïwan tel que pratiqué dans les années 1990 (le Sinica Treebank), et d'autre part d'échantillons d'un corpus historique divisé en trois périodes distinctes (chinois archaïque 上古, chinois médiéval 中古 et chinois pré-moderne 近世).

Le corpus du Shun-pao est un corpus inédit, dont l'accès au texte intégral n'a pu être rendu possible que tout récemment grâce au projet ERC *Elites, Networks and Power in Modern China*. Il contient le

texte de tous les articles (en excluant les publicités) qui ont été OCRisés et corrigés manuellement. Il comporte environ 750 millions de sinogrammes. Une version des données avec découpage en pages et en articles a pu être obtenu dans un second temps, ce qui a fait disparaître certaines limites des travaux précédents. Mais le découpage en articles nous semble encore très imparfait. La typographie est aussi en grande partie perdue, ce qui a posé problème pour l'étude de l'évolution des pratiques de ponctuation, mais n'est pas limitant pour la présente étude.

3 L'approche par « modèles de langue »

Des travaux précédents (Magistry, 2019) à la suite desquels nous nous situons reposent sur l'entraînement de modèles langues à partir du texte brut des documents. Notons que cette stratégie était en partie imposée par l'absence de métadonnées, par une segmentation limité du corpus et par l'absence de points de comparaison avec les autres corpus, invoqués dans un second temps.

Avec une segmentation par année du corpus, il est tout de même possible de proposer une périodisation basée sur la perplexité d'une multitude de modèles de langue (un par année) à la lecture de chaque autre année du corpus. On peut ainsi construire une matrice de distances et appliquer du clustering hiérarchique pour repérer des points de rupture.

De plus, en utilisant des plongements lexicaux contextualisés obtenus par l'entraînement de modèles de langues avec des BiLSTM (Akbik et al., 2018), on a pu suivre l'évolution de l'usage de certains items lexicaux choisis dans la littérature de linguistique historique (Peyraube, 1996; Coblin, 2000)

4 L'analyse des correspondances

Après l'obtention de corpus comparables et d'une segmentation du corpus plus fine (a minima à la page), on peut recourir à l'analyse factorielle des correspondances (Benzécri, 1973). On dispose ainsi d'une méthode éprouvée pour situer la langue et visualiser son évolution en se basant sur des comptes d'occurrences de mots « vides » (ou grammaticaux), que l'on estime plus caractéristiques de la langue d'un texte que les mots « pleins » qui peuvent plus facilement être des emprunts ou être surtout caractéristiques des thèmes et du genre d'un texte. On peut ainsi obtenir des plans factoriels qui situent des pages de notre corpus par rapport aux chinois anciens et modernes.

5 Points étudiés

On étudie en particulier certains facteurs et propriétés des modèles pour dégager leurs atouts respectifs. Il s'agit des propriétés des corpus utilisés en taille, en qualité des annotations, en finesse des métadonnées et des objets mis au centre des analyses (ensemble de textes, document unique, type ou occurrence token). On observe aussi des comportements différents face aux variantes graphiques des sinogrammes, qui n'ont été standardisés que récemment et au problème de la segmentation en mots.

On montrera qu'en combinant ces méthodes, il envisageable d'enrichir notre corpus en catégorisant les textes qui le composent et aussi de suivre des évolution diachronique de l'emploi de certains items lexicaux.

Références

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Benzécri, J.-P. (1973). *L'analyse des données 2 l'analyse des correspondances*. #0, Dunod, Paris, Bruxelles, Montréal.

Coblin, W. S. (2000). A brief history of mandarin. *Journal of the American Oriental Society*, 120(4) :537–552.

Magistry, P. (2019). Languages(s) of the SHUN-PAO, a Computational Linguistics account. In *10th International Conference of Digital Archives and Digital Humanities*, Taipei, Taiwan.

Peyraube, A. (1996). Recent issues in chinese historical syntax. In Huang, C.-T. J. and Li, Y.-h. A., editors, *New Horizons in Chinese Linguistics*, pages 161–213. Springer Netherlands, Dordrecht.

On avertives as complex negative event descriptions

So-called 'avertive' and 'frustrative' inflectional markers were (to the best of my knowledge) first identified in non-Pama-Nyungan Australian languages, notably in Rembarnga (non-Pama-Nyungan/Maningrida) (McKay 1975) for and for Iwaidja (non-Pama-Nyungan/Iwaidjan); their meaning can be best paraphrased as 'I nearly V-ed/was going to V-ed [but didn't]', cf. (1).

(1) maju ngan-ambija-na
WANT.Part 1sg.FRUST-laugh-FRUST
'I was going to laugh (but I didn't')' (Iwaidja; (Pym & Larrimore 1979: 76))

Amazonian languages have been diagnosed as another major linguistic area where grammatical avertive or frustrative markers are abundant, cf. e.g. (Aikhenvald 2012: 185); (Overall 2017) formulated the first in-depth, areal-typological study of such markers in Amazonia. And while (Kuteva 1998; Kuteva et al. 2015) formulated a foundational account of avertives and frustratives in terms of grammaticalization theory, and offered wider typological glimpses into the actual extent of these categories (see also (Alexandrova 2016)), the study of avertive and frustrative markers is still in its infancy. Indeed, they seem to be just as widespread across languages of the world as other recently uncovered categories, such as e.g. evidentiality, and is far from being well described (let alone understood) at the present time.

This talk will offer an initial areal-typological account of the complex semantics of avertive and frustrative markers in the North of Australia, beginning with a description of Iwaidja and Anindilyakwa (non-Pama-Nyungan/Gunwinyguan) avertives/frustratives, followed by a short typological study of similar markers in a sample of Arnhem Land, non-Pama-Nyungan languages (nPN). What renders nPN avertives/frustratives unique and especially worthy of consideration is their striking polyfunctionality, cf. (2), as they are used to convey a variety of meanings comprising:

- Negative past events
- Positive or negative past modals (esp. deontic/volitional)
- And positive or negative past imperfective/proximative meanings

(2) karlu ayana-wu-ni (iwaidja)
neg 1sg>3pl.FRUST-hit-FRUST (TAIM20181114DY@00:04:11)
1. 'I didn't hit them.'
2. 'I should not have hit them'.
3. 'I wasn't going to hit them'. [= 'it was predictable I wasn't going to hit them']

I will suggest these three readings cannot be productively derived from a single, underspecified meaning, either by pragmatic or semantic means – in particular, I will hypothesize that the negative past event meaning of e.g. the Iwaidja frustrative is non-productive, i.e. cannot be compositionally derived from any modal base (in the sense of (Kratzer 1991)) or conversational background (in the sense of (Kratzer 2012)) lexically encoded by the corresponding inflection. Instead, I will suggest it is the byproduct of the semanticization of a former conventionalized implicature (CI). In order to clarify this idea, I will sketch a formal implementation for the various meanings listed above; while the past counterfactual negative deontic meaning (reading #2) and the negative volitional meaning (reading #3) of (2) will be accounted for compositionally, its negative past event reading (#1) will be represented using (Gutzmann 2011)'s extension of the multidimensional semantics of (Potts 2005; Potts 2007): more specifically, I will propose that the negative past event meaning of (2) was diachronically derived from a (possible) contextual implicature associated with proximative meaning #3 ('it was predictable I wasn't going to V' > I didn't V'), later promoted to the status of CI – and from there, to the statue that of at-issue meaning.

Last but not least, I will argue that not all avertive inflections have equally developed semanticized negative meanings, and that the typology of avertivity in Australia should offer a complex and nuanced picture of because of avertives at the morphosyntax to semantics and pragmatics interface, with some forms merely *implicating*, and others *conveying* negative past events. Only the latter effectively contribute complex event descriptions, combining a *bona fide* negative past event, with an underlying/secondary past modal event.

Selected references

Aikhenvald, Alexandra Y. 2012. *The Languages of the Amazon*. Oxford: OUP.

Alexandrova, Anna. 2016. Avertive constructions in Europe and North Asia : An areal typology. Conference talk presented at the Chronos 12 - 12th International Conference on Actionality, Tense, Aspect, Modality/Evidentiality, Université de Caen.

Gutzmann, Daniel. 2011. Expressive Modifiers & Mixed Expressives. In O. Bonami & P. Cabredo Hofherr (eds.), *Empirical Issues in Syntax and Semantics 8*, 123–142. Université Paris-Sorbonne.

Haspelmath, Martin. 2001. The European linguistic area: Standard Average European. In M. Haspelmath, E. König, W. Oesterreicher & W. Dressler (eds.), *Language Typology and Language Universals / Sprachtypologie und sprachliche Universalien / La typologie des langues et les universaux linguistiques. Volume 2*, 1492–1510. Berlin: Walter de Gruyter.

Kratzer, Angelika. 1991. Modality. In A. von Stechow & D. Wunderlich (eds.), *Semantics: An International Handbook of Contemporary Research*, 639–650. Berlin: De Gruyter.

Kratzer, Angelika. 2012. *Modals and Conditionals: New and Revised Perspectives*. Oxford: Oxford University Press.

Kuteva, Tania A. 1998. On Identifying an Evasive Gram: Action Narrowly Averted. *Studies in Language* 22(1). 113–160.

Kuteva, Tania, Bas Aarts, Gergana Popova & Anvita Abbi. 2015. On five counter-to-fact grammatical categories. Presented at the Conference "Diversity Linguistics - Retrospect and Prospect", Max-Planck-Gesellschaft, Leipzig.

McKay, Graham Richard. 1975. *Rembarnga : a language of central Arnhem Land*. Canberra: Australian National University Ph.D. Thesis.

Overall, Simon E. 2017. A Typology of Frustrative Marking in Amazonian Languages. In A. Y. Aikhenvald & R. M. W. Dixon (eds.), *The Cambridge Handbook of Linguistic Typology*, 477–512. Cambridge: CUP.

Potts, Christopher. 2005. *The Logic of Conventional Implicatures*. Oxford / New York: OUP.

Potts, Christopher. 2007. The expressive dimension. *Theoretical Linguistics* 33(2). 165–198.

Pym, Noreen & Bonnie Larrimore. 1979. *Papers on Iwaidja Phonology and Grammar* (Work Papers SIL-AAB, Series A). Darwin: Summer Institute of Linguistics.

Quelques exemples de négation dans les langues créoles

Emmanuel Schang
LLL (UMR 7270), 10 rue de Tours, BP 46527, 45072 Orléans Cedex 2, France
emmanuel.schang@univ-orleans.fr

MOTS-CLÉS : Négation, créoles, concordance négative.

KEYWORDS: Negation, negative concord, creole languages.

1 Introduction

De nombreuses monographies proposent des études formelles très poussées sur la négation. (Chierchia, 2013) est une illustration de la richesse des analyses possibles sur cette langue extrêmement bien connue qu'est l'anglais. On dispose cependant rarement d'études aussi fouillées sur les langues qui ne font pas partie des langues romanes ou des langues germaniques.

(Déprez and Henri, 2018) note dans son introduction que les langues créoles sont ignorées dans les travaux théoriques portant sur la concordance négative, tels que (Zeijlstra, 2004) et (Haegeman and Zanuttini, 1996) par exemple. Si (Déprez and Henri, 2018) comble en partie cette lacune en dédiant 12 chapitres à la description de la négation dans plusieurs langues créoles, beaucoup reste encore à être mis au jour sur ce point. En effet, les langues créoles sont réputées 'simples' (comme le prétend (McWhorter, 2011)) et l'Atlas of Pidgin and Creole Structures (APICS https ://apics-online.info/) ne rend pas justice à la complexité des faits constatés. De fait, des éléments potentiellement intéressants pour toute théorie de la négation ayant une prétention d'universalité mériteraient plus d'attention.

Cet article a pour modeste objectif de mettre à disposition quelques données sur le fonctionnement de la négation en gwadloupéyen, en santomense et en principense.

2 La négation dans deux créoles à base lexicale portugaise : le santomense et le principense

Le santomense (ou forro, noté ST) est un créole à base lexicale portugaise parlé sur l'île de São Tomé (Golfe de Guinée). Cela signifie que son lexique est en majeure partie issue du portugais. Mais sa grammaire en revanche est significativement différente du portugais (que ce soit la variété du Portugal ou du Brésil). Le principense (PR), aussi appelé lung'Ie (littéralement : la langue de l'île), est un autre créole portugais, parlé sur l'île de Príncipe. Les deux îles sont distantes mais appartiennent toutes deux à la République Démocratique de São Tomé et Príncipe.

A la différence du portugais, le ST fait usage d'une négation double, comme le montre (1).

(1) a. Ele não comprou arroz. [PT]
3SG NEG acheter riz
'Il n'a pas acheté de riz'
b. ê na kompla loso fa [ST]
3SG NEG acheter riz NEG
'Il n'a pas acheté de riz'

On a donc :

$$\text{P : não} \Longrightarrow \neg$$

$$\text{ST : na ... fa} \Longrightarrow \neg$$

Cependant, seule la négation préverbale est maintenue dans les propositions exprimant la cause (2-a) et seule la négation finale apparaît dans les contextes contrastifs (2-b) :

(2) a. ê fuji pa Tataluga na da ê sotxi.
3SG fuire pour Tortue NEG donne 3SG fouet
'Il s'est enfui pour que Tortue ne le batte pas'
b. ami fa
1SG NEG
'Pas moi.'

Contrairement au ST, le principense n'utilise principalement qu'une négation finale *fa*. (Maurer, 2009) détaille cependant des usages dans lesquels *na* préverbal est utilisé. La différence de fonctionnement entre le ST et le PR, qui sont deux créoles liés historiquement, a fait l'objet de plusieurs études ((Hagemeijer, 2009; Güldemann and Hagemeijer, 2019)).

Nous rappellerons ici les particularités du PR concernant la négation.

3 La négation en gwadloupéyen

Le fonctionnement de la négation et des différents mots négatifs (concordance négative) a fait l'objet d'une étude détaillée dans (Petitjean and Schang, 2018). Nous reprendrons ici les points principaux, notamment :

— la négation et ses modifications morphologiques au contact de l'adverbe *anko* 'encore' et de la marque du futur *ké*,
— les mots négatifs et la concordance négative.

4 Conclusion

Ce travail expose au regard des linguistes formalistes et informaticiens des points assez peu connus qui illustrent la complexité du fonctionnement de la négation, tant synchroniquement que diachroniquement, dans quelques langues créoles.

Références

Chierchia, G. (2013). *Logic in grammar : Polarity, free choice, and intervention*. OUP Oxford.

Déprez, V. and Henri, F. (2018). *Negation and Negative Concord : The View from Creoles*, volume 55. John Benjamins Publishing Company.

Güldemann, T. and Hagemeijer, T. (2019). The history of sentence negation in the gulf of guinea creoles. *Journal of Ibero-Romance Creoles*, 9(1) :55–84.

Haegeman, L. and Zanuttini, R. (1996). Negative concord in west flemish. *Parameters and functional heads. Essays in comparative syntax*, pages 117–197.

Hagemeijer, T. (2009). Aspects of discontinuous negation in santome. *Negation patterns in West African languages and beyond*, pages 139–165.

Maurer, P. (2009). *Principense-Grammar, texts, and vocabulary of the Afro-Portuguese creole of the island of Príncipe, Gulf of Guinea*. London-Colombo : Battlebridge Publications.

McWhorter, J. H. (2011). *Linguistic simplicity and complexity : Why do languages undress ?*, volume 1. Walter de Gruyter.

Petitjean, S. and Schang, E. (2018). Sentential negation and negative words in guadeloupean creole. In Déprez, Viviane and Henri, Fabiola, editor, *Negation and Negative Concord : The View from Creoles*.

Zeijlstra, H. (2004). *Sentential negation and negative concord*. PhD thesis, Netherlands Graduate School of Linguistics.

Reading interlinearized glossed texts: inference of linguistic features from free translations

Sylvain Loiseau[1]
(1) UMR 7107 Lacito & Université Sorbonne Paris Nord
sylvain.loiseau@univ-paris13.fr

MOTS-CLÉS : IGT (textes interlinéarisés), divergence structurelles, traduction.

KEYWORDS: IGT (interlinearized glossed text), structural divergences, translation.

A large set of texts representing various languages of the world have been edited and published by linguists working on linguistic diversity. Several conventions have been used, but one is particularly common, the notation known as "interlinear glossed text" (IGT) (Lehmann, 2004).

(1) Kolyma Yukaghir (Maslova, 2003, 184)

irk-in	čas-ek	tet-in	kej-nu-me
one-ATTR	hour-PRED	you-DAT	give-IPFV-OF :1SG

'I give you one hour'

Rather than a formal scheme it is a set of loosely converging practices that have varied widely over time, as well as according to theoretical persuations or research interests (Bow et al., 2003; Pasquereau, 2010). However, a trend towards standardization of IGT has emerged in recent years (Bickel et al., 2008). In its various guises, a set of fundamental characteristics of IGT can be identified :

— it contains at least three tiers (lines) : transcription, morphem gloss, free translation.
— tier 1 give the object language, tier 2 and 3 are labelled in a meta language
— tier 1 and 2 are aligned at word and morpheme levels

In recent years, a shared interest among linguists and computational scientists has emerged for better automatic processing of IGTs. More linguistically diverse data can present interesting challenges and offer better formal generalisations, while computational tools can help construct typological generalizations.

IGT have been studied by NLP in order to retreive, parse and aggretate them into corpora (Lewis and Xia, 2010). Most often, the goal is to produce automatic glosses of unglossed texts through machine learning of relations between tier 1 and tier 2 (Zhao et al., 2020; McMillan-Major, 2020).

IGTs are mostly treated in such avenue of research as if they were tagged corpora. The third tier, the free translation, is not taken into account. However, this format is not intended to represent a machine-readable data structure, it is a text intended for human readers. The reader is expected to do all kind of inferences, in particular using the free translation tier. This would include, for instance :

— the actual acception of lexems. It is a rule of IGT that a lexem, even polysemic, should be

always render with the same basic/generic translation in the gloss, while the actual acception is given in the gloss (Lehmann, 2004, 8) (see (a) bellow : 'hear' in the gloss, 'understand' as the actual acception in that context, in the free translation) ;
— identification of multi-word expressions (the gloss tier gives literal meanings of each morphems, the third line is key to guess that a group of words combines into MWE) In (b) bellow, a conventionalized serial verb construction implying 'walk' and 'go' means 'make a walk' ;
— idiomatic meaning of the sentence as a whole cannot be asserted without the third line (see (c) bellow) ;
— the free translation may provide necessary information for the grouping of tokens of tiers 1 (and 2) into constituents. Grammatical relations (if not overtly marked by the morphology in the gloss tier), as well as semantic roles of arguments, are also infered from the translation by projecting syntactic structure from tier 3 to tier 2 and 1.

(2) a. Cavineña (Guillaume, 2008, 787)

era baka-aje-nuka-ya=dya
1SG-ERG hear-GO.DIST-REITR-IMPFV=FOC

'I managed to understand (lit. hear) it.'

b. Usan (Reesink, 1987, 74)

wuri ibeibi gib ig-our
they walking go.SS be-3P.PR

'They are walking around [=out for a walk, purposeless]'

c. Murinjpata (Walsh, 1976, 407)

ŋaɻa-ɻe t̪e-t-aŋd̪in
what-INSTRUM ear-2SG-have

'how do you know (Literally : with what/by what means do you have an ear ?)'

An IGT represents linguistic structures not only through the representation of the morphological structure in tier 2, but also through the syntactic and semantic structures of tier 3. This mapping is largely informal, and very complex to make fully explicit, since it is a matter of translation skills. Each line of an IGT contribute to the understanding of the linguistic structures of the sentence. In (Xia and Lewis, 2007), the syntactic tree of the free translation is projected on the object language sentence. However, the free translation is not complementing the glosses with syntactic information only. It is also providing key information for the interpretation of lexical and semantic structures.

A systematic inventory and a typology of the inferences that are made by the reader of IGT when comparing the structures of the meta-language (used in the free translation) and the structures of the object language is a necessary step toward making explicit the pieces of information that are present in the IGT but in an informal manner, and then toward making IGT maximally usefull for computer-assisted linguistic research more generaly. In order to model these inferences, it may be usefull to take into account the structural divergences between languages pairs such as configurationality, referential density (Bickel, 2003), information packaging strategies, alignment or morphological type.

Références

Bickel, B. (2003). Referential density in discourse and syntactic typology. *Language*, 79(4) :708–736.

Bickel, B., Comrie, B., and Haspelmath, M. (2008). The leipzig glossing rules. conventions for interlinear morpheme by morpheme glosses. *Revised version of February*.

Bow, C., Hughes, B., and Bird, S. (2003). Toward a general model of interlinear text. In *Proceedings of Emeld workshop 2003*.

Guillaume, A. (2008). *A Grammar of Cavineña*, volume 44 of *Mouton Grammar Library*. Mouton de Gruyter.

Lehmann, C. (2004). Interlinear morphemic glosses. In Booij, G. E., Lehmann, C., Mugdan, J., and Skopeteas, S., editors, *Morphologie : Ein Internationales Handbuch*, volume 2 of *Handbücher zur Sprach- und Kommunikationswissenschaft (HSK)*, pages 1834–1857. Walter de Gruyter, Berlin.

Lewis, W. and Xia, F. (2010). Developing odin : A multilingual repository of annotated language data for hundreds of the world's languages. *Journal of Literary and Linguistic Computing*, 25(3) :303–319.

Maslova, E. (2003). *A grammar of Kolyma Yukaghir*. Mouton Grammar Library. Mouton de Gruyter, Berlin ; New York.

McMillan-Major, A. (2020). Automating gloss generation in interlinear glossed text. In *Proceedings of the Society for Computation in Linguistics*, volume 3(1), pages 338–349.

Pasquereau, J. (2010). Les pratiques de gloses en français, espagnol et anglais. Master's thesis, Lyon 2, Lyon.

Reesink, G. P. (1987). *Structures and their functions in Usan*. Studies in Language Companion Series. John Benjamins Publishing Company.

Walsh, M. (1976). Murinjpata. In Dixon, R., editor, *Grammatical categories in Australian Languages*, number 22 in Linguistic Series, pages 405–408. Australian Institute of Aboriginal Studies / Humanities Press Inc, New Jersey USA / Canberra.

Xia, F. and Lewis, W. D. (2007). Multilingual structural projection across interlinear text. In *Proceedings of NAACL HLT 2007*, pages 452–459.

Zhao, X., Ozaki, S., Anastasopoulos, A., Neubig, G., and Levin, L. (2020). Automatic interlinear glossing for under-resourced languages leveraging translations. In *International Conference on Computational Linguistics (COLING)*.

Segmentation en mots faiblement supervisée pour la documentation automatique des langues

Shu Okabe[1] François Yvon[1] Laurent Besacier[2]
(1) Université Paris-Saclay, CNRS, LISN, Bât. 508, Rue du Belvédère, F-91405 Orsay, France
(2) NAVER LABS Europe, 6-8 chemin de Maupertuis, F-38240 Meylan, France
shu.okabe@limsi.fr, francois.yvon@limsi.fr,
laurent.besacier@naverlabs.com

MOTS-CLÉS : segmentation en mots, documentation automatique des langues, modèle bayésien non paramétrique.

KEYWORDS: word segmentation, computational language documentation, Bayesian non-parametric model.

1 Introduction

La documentation automatique des langues vise à outiller les linguistes de terrain pour faciliter l'annotation des données linguistiques. Les travaux récents se sont concentrés sur des méthodes non-supervisées. Toutefois, comme le souligne Bird (2020), des ressources auxiliaires, par exemple des listes de mots ou des textes annotés, sont souvent disponibles grâce aux efforts passés et présents de locuteurs natifs et de linguistes. L'idée de cette étude est de mobiliser ces ressources dans les algorithmes de segmentation en mots pour améliorer leurs performances.

La tâche de segmentation en mots consiste à retrouver les frontières des mots dans une séquence continue de caractères. Elle intervient notamment en documentation automatique des langues quand des enregistrements audio sont retranscrits phonétiquement. Une illustration de cette tâche est dans (Godard, 2019), qui étudie différents modèles de segmentation notamment des modèles bayésiens non paramétriques sur une langue peu dotée : le Mboshi (Bantu C25). Dans ce contexte, nous étudions comment des ressources auxiliaires peuvent aider les modèles de segmentation en mots.

2 Méthodes

2.1 Conditions expérimentales

Modèle Le modèle de référence utilisé est la version unigramme de `dpseg` (Goldwater et al., 2009), un modèle simple et bien adapté au traitement de petits corpus. Ce modèle repose sur les processus de Dirichlet pour calculer la probabilité de mots et de séquences de mots. Dans notre implémentation, l'inférence de ce modèle repose sur l'échantillonnage de Gibbs et le recuit simulé. Une variante de ce modèle, basée sur les processus de Pitman-Yor (PYP), a aussi été implémentée (Teh, 2006).

Supervision faible Deux types de ressources ont été considérées pour introduire une supervision faible, qui simulent des conditions réelles de documentation : des phrases partiellement ou complètement segmentées et des listes de mots. La première situation correspond au repérage de pauses dans les enregistrements (annotation partielle) ou à des phrases déjà segmentées (annotation complète). La seconde correspond à la pré-existence de dictionnaires ou au recueil des types observés dans les phrases segmentées. Ces listes sont utilisées pour renforcer la probabilité à priori des types connus.

Langues étudiées Deux langues peu dotées en cours de documentation ont été étudiées : le Mboshi, langue bantoue déjà présentée dans des travaux antérieurs (Godard et al., 2018), ainsi que le Japhug, langue sino-tibétaine parlée dans la partie ouest de la Chine (Jacques, 2021).

2.2 Expériences

Stratégie de supervision faible
La première expérience évalue l'impact d'une supervision faible sur les deux modèles (`dpseg` et sa variante utilisant PYP). Lorsque l'on supervise les frontières de mots, les annotations denses sont plus efficaces que des annotations partielles réparties aléatoirement dans le texte. Dans le cas de listes de mots, l'amélioration du modèle de caractères, associée à une augmentation de la probabilité des mots présents dans le dictionnaire de supervision, permettent d'obtenir les meilleurs résultats.

Le tableau 1 détaille ces résultats pour le Mboshi, avec les modèles `dpseg` et `pypseg`, sans supervision (/), avec une supervision sur les occurrences (phrases annotées) et avec une supervision sur les types (listes de mots). Les données de supervision sont dérivées d'une annotation de 200 phrases du corpus d'entraînement. Les segmentations sont évaluées avec les F-scores à trois niveaux : BF pour l'évaluation des frontières de mots, WF pour l'évaluation au niveau des occurrences et LF pour l'évaluation au niveau des types.

modèle	dpseg			pypseg		
supervision	/	token	type	/	token	type
BF	65.9	68.7	66.4	66.2	68.8	65.8
WF	37.6	42.4	39.4	37.9	42.5	38.7
LF	23.8	31.4	40.0	24.5	31.6	39.9

TABLE 1 – Résultats des segmentations non supervisées et supervisées sur le texte Mboshi

Pour les deux modèles, la supervision faible améliore la segmentation, comme en témoigne la hausse notable des scores pour toutes les métriques, à l'exception du BF dans le cas `pypseg` avec une supervision par liste de mots.

Dans le cas non supervisé, le modèle reposant sur les PYP apparaît meilleur que `dpseg`. Toutefois, dans les deux autres cas, il ne semble pas améliorer de manière significative les performances des modèles faiblement supervisés en comparaison avec leurs versions `dpseg`.

Dans l'ensemble, le modèle `dpseg` supervisé avec un dictionnaire de mots obtient le meilleur résultat.

Apprentissage incrémental
La seconde expérience étudie un scénario d'apprentissage incrémental et simule la situation où un

expert annote progressivement des phrases. Les corrections sont prises en compte pour l'annotation des phrases ultérieures (`regular`). Une variante a aussi été implémentée, dans laquelle le modèle probabiliste de base qui évalue la forme des unités lexicales dans le modèle est régulièrement mis à jour en intégrant les mots au fur et à mesure de leur vérification (`2level`). Cette seconde approche a présenté de meilleurs résultats, avec un effet stable à travers tout le texte.

Le graphique 1 présente l'évolution du taux d'erreur tout au long du texte en Japhug pour le modèle de référence ainsi que les deux modèles faiblement supervisés. Le taux d'erreur a été calculé toutes les 100 phrases, en divisant le nombre d'erreurs par la longueur de ces phrases. Le point de départ de ces trois modèles est la sortie du modèle `dpseg` complètement non supervisé. Régulièrement, le modèle effectue des itérations supplémentaires d'échantillonnage de Gibbs sur l'ensemble du texte restant afin de propager les améliorations obtenues grâce aux phrases corrigées.

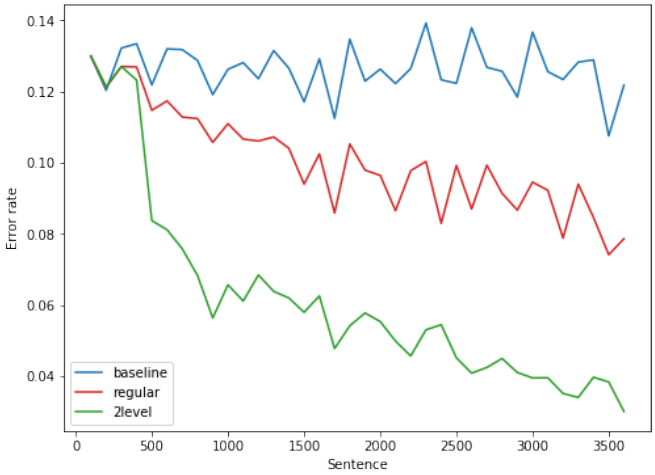


FIGURE 1 – Résultats de l'apprentissage incrémental pour le texte Japhug 3K

Le modèle de référence (en bleu) maintient un taux d'erreur moyen à peu près constant sur tout le texte. À l'inverse, les deux modèles qui bénéficient de l'apprentissage en ligne permettent une chute tendancielle du taux d'erreur ; le modèle `2level` (en vert) se montre sensiblement meilleur, grâce à son modèle lexical qui devient de plus en plus précis.

Comparaison de segmentation en mots ou en morphème

Une dernière expérience porte sur l'étude des segmentations en mots et morphèmes.

référence	mot			morphème		
supervision	/	mot	morph.	/	mot	morph.
BF	72.9	78.8	76.1	80.8	71.0	75.8
WF	45.7	55.8	51.0	54.7	39.2	45.1
LF	20.1	42.7	32.8	41.2	33.5	43.8

TABLE 2 – Comparaison des résultats sur le texte Japhug 3K pour un texte segmenté en mot ou en morphème (référence), avec ou sans supervision sur les types (supervision)

Le tableau 2 présente les résultats comparatifs pour cette expérience qui utilise comme supervision, soit un dictionnaire de mots, soit un dictionnaire de morphèmes (extrait de 200 phrases dans les deux cas). Sans supervision, les modèles comparés ont tendance à produire une segmentation en unités

courtes, proche d'une segmentation en morphèmes ; ajouter une supervision par dictionnaire de mots permet de contre-balancer cette tendance (de manière plus mitigée lors d'une supervision sur les morphèmes).

3 Conclusion

Plusieurs stratégies de supervision faible ont été étudiées pour la segmentation de mots de langues peu dotées. Dans le cadre de la documentation automatique de langues, des ressources auxiliaires, telles des phrases annotées ou des listes de mots, sont souvent disponibles et peuvent être mobilisées pour pallier la faible quantité de données. Les modèles bayésiens non paramétriques parviennent à bénéficier de ces données supplémentaires pour les deux types de ressources. L'apprentissage incrémental semble aussi se prêter à une utilisation réelle, tandis que l'expérience de segmentation à deux niveaux, mots et morphèmes, ouvre des perspectives pour de futurs travaux.

Remerciements

Ce travail est effectué dans le cadre du projet franco-allemand « La documentation automatique des langues à l'horizon 2025 » (*Computational Language Documentation by 2025*, CLD 2025, ANR-19-CE38-0015-04) . Un rapport détaillé de l'ensemble de ces expériences sera mis en ligne sur le site du projet. Les auteurs remercient Guillaume Jacques pour la mise à disposition des textes annotés en Japhug.

Références

Bird, S. (2020). Decolonising Speech and Language Technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Godard, P. (2019). *Unsupervised word discovery for computational language documentation*. Theses, Université Paris-Saclay.

Godard, P., Adda, G., Adda-Decker, M., Benjumea, J., Besacier, L., Cooper-Leavitt, J., Kouarata, G.-N., Lamel, L., Maynard, H., Mueller, M., Rialland, A., Stueker, S., Yvon, F., and Zanon-Boito, M. (2018). A very low resource language speech corpus for computational language documentation experiments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Goldwater, S., Griffiths, T. L., and Johnson, M. (2009). A Bayesian framework for word segmentation : Exploring the effects of context. *Cognition*, 112(1) :21–54.

Jacques, G. (2021). *A grammar of Japhug*. Language Science Press.

Teh, Y. W. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 985–992, Sydney, Australia. Association for Computational Linguistics.

Simplification syntaxique de textes à base de représentations sémantiques en DMRS

Rita Hijazi[1,2]

(1) LPL, 5 av. Pasteur, Aix-en-Provence, France
(2) LIS, 52 Av. Escadrille Normandie Niémen, Marseille, France
rita.hijazi@lpl-aix.fr

MOTS-CLÉS : Simplification Automatique de Textes, simplification syntaxique, réécriture de graphes, représentations sémantiques à base de graphes, DMRS.

KEYWORDS: Automatic text simplification, graph-based meaning representations, graph rewriting, DMRS.

1 La simplification automatique de textes

La simplification de textes (SAT) est un domaine du traitement automatique des langues (TAL) visant à rendre des textes plus abordables tout en garantissant l'intégrité de leur contenu. Saggion (2017) définit la SAT comme étant le processus de transformation d'un texte en un autre texte qui véhicule le même contenu sémantique, afin de le rendre plus facile à lire et à comprendre par un public cible. Dans notre travail, nous nous limitons à la problématique de la simplification syntaxique (SS) ayant pour but d'identifier et de transformer des phrases contenant des constructions syntaxiques qui peuvent nuire à la lisibilité et à la compréhension pour certaines personnes en équivalents plus lisibles ou compréhensibles.

2 De l'usage de la sémantique pour la simplification syntaxique de textes

Notre but est de proposer une méthode de SS basée sur une analyse sémantique et un système permettant de la mettre en œuvre de façon automatique. Cette méthode et ce système doivent permettre la simplification syntaxique de textes par le traitement de constructions syntaxiques complexes, telles que les subordinations, les constructions appositives, les coordinations et la double négation.

En simplification syntaxique, décider quand et où diviser une phrase en deux phrases plus simples, associées à des évènements distincts, peut être déterminé par la syntaxe même, par exemple, des phrases avec des subordonnées. Des approches de la simplification syntaxique essentiellement basées sur la syntaxe (Zhu et al., 2010 ; Woodsend et Lapata, 2011) n'arrivent souvent pas à reconstruire correctement l'élément partagé ni à identifier correctement le point de découpage. En effet les représentations sémantiques des phrases donnent une idée claire des événements, de leurs

ensembles de rôles associés et des éléments partagés, facilitant ainsi à la fois l'identification des points de découpage possibles et la reconstruction des éléments partagés dans les phrases résultant du découpage (Narayan et Gardent, 2014 ; 2016).

Notre recherche porte sur la proposition d'une méthode et d'un système associé de SS de phrases. Cette simplification est basée sur des transformations de représentation sémantique des phrases à simplifier, transformations réalisées par des règles définies manuellement.

L'usage de la sémantique uniquement n'est pas suffisant pour détecter les constructions syntaxiques comme les appositives et les subordinations. C'est pourquoi nous nous basons sur un formalisme sémantique qui offre des informations sur les rôles thématiques nécessaires pour la reconstruction de l'élément partagé d'une part, et d'autre part donnant des informations syntaxiques pour détecter les constructions syntaxiques complexes. Nous avons retenu une représentation sémantique de texte spécifique, une représentation sémantique profonde, la notation DMRS (*Dependency Minimal Recursion Semantics*) attribuée par Copestake (2009).

Cette notation DMRS prend en compte à la fois les annotations sémantiques et syntaxiques des phrases. Cela permet un compte rendu linguistique de l'opération de division en ce que les constructions complexes sont exprimées dans les représentations et les éléments partagés sémantiquement sont prises en compte afin de les réécrire dans la phrase découpée. Ce qui entraine à avoir une sortie significativement plus simple qui est à la fois grammaticale (informations syntaxiques du DMRS) et préservant le sens (informations liées à la sémantique dans DMRS). DMRS offre des informations sur les rôles thématiques nécessaires pour la reconstruction de l'élément partagé d'une part, et d'autre part donnant des informations syntaxiques pour détecter les constructions syntaxiques complexes.

2.1 Méthode proposée

Notre recherche concerne essentiellement la langue anglaise. Nous avons retenu une approche à base de règles construites manuellement. Cette approche requière d'implication humaine pour la définition des règles, mais conduit à des systèmes de simplification précis. Selon Siddharthan (2014), les règles manuelles sont utilisées dans le domaine de la simplification du texte lorsqu'un système se concentre sur des structures et des phénomènes linguistiques très spécifiques qui sont relativement faciles à gérer avec un ensemble limité de règles.

La méthode de simplification syntaxique proposée repose sur trois étapes principales. La première étape consiste à représenter la phrase complexe par un graphe DMRS. Dans la seconde étape, il s'agit de transformer ce graphe DMRS en un ou plusieurs autres graphes DMRS en appliquant un ensemble de règles de transformation. Enfin, dans la troisième étape, il s'agit de générer à partir de ces nouveaux graphes obtenus les phrases simplifiées associées à la phrase complexe de départ. La densité propositionnelle du texte est ainsi réduite, ce qui conduit à une meilleure compréhension de chacune de ses phrases. La division est basée sur la préservation d'un ordre SVO (sujet-verbe-objet) dans chaque division, qui pourrait être considéré comme une structure agent-verbe-patient.

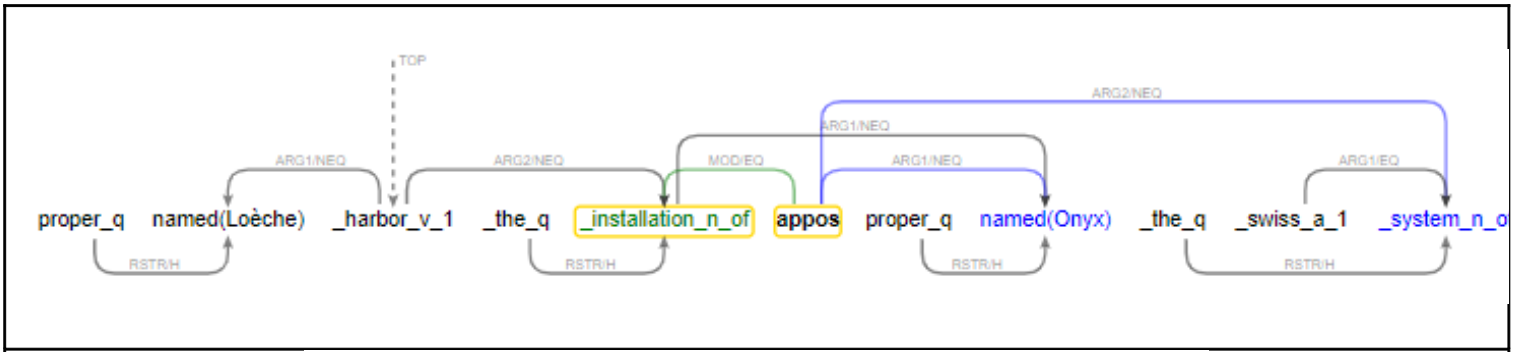
2.2 Système de simplification syntaxique associé

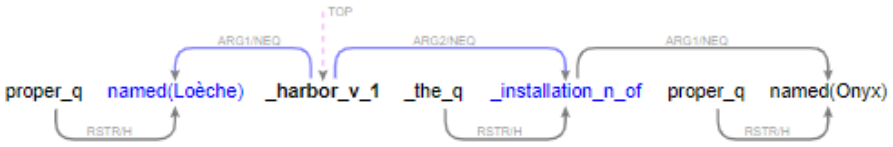
Ce système de simplification syntaxique automatique met en œuvre la méthode de simplification syntaxique précédente. Il est basé sur un ensemble de modules logiciels existants qui sont intégrés dans un « pipe » logiciel spécifique. Le premier module logiciel est un analyseur (*parser*) DMRS (ACE)[1] permettant de construire la représentation de la phrase en un graphe, le module logiciel GREW (*Graph REWriting system* ; Guillaume et al., 2012), un système de réécriture de graphes permettant la transformation de ce graphe DMRS en une ou plusieurs autres graphes DMRS selon des règles de transformation que j'aurai définies par ailleurs. Enfin, la mise en œuvre d'un générateur (ACE toujours) permettant la génération des phrases simplifiées.

3 Exemple d'application : le traitement des appositives

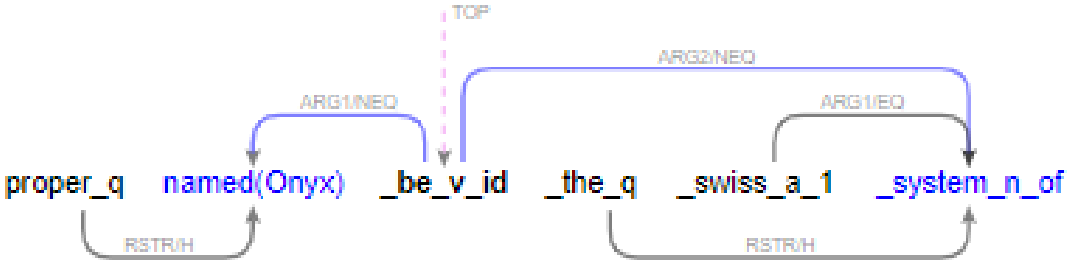
En DMRS, l'apposition est formée de deux noms adjacents décrivant la même référence dans une phrase. Elle peut être capturée avec précision : elle est identifiée par une relation de type *appos* qui prend les deux noms adjacents comme arguments (exemple 1). La figure 1 est la représentation DMRS de la phrase. Les figures 1a et 1b est la représentation DMRS des deux sous-phrases.
La règle de découpage d'apposition supprime d'abord le nœud *appos* ainsi que le prédicat (*harbor*) et son ARG2 pour former une première DMRS, puis il construit l'autre DMRS en remplaçant l'ARG1 d'*appos* par son ARG2 et en supprimant ensuite *appos* et son ARG1.

(1) Loèche harbours the installations of Onyx, the Swiss system.





1.a. DMRS de la phrase « Loèche harbours the installations of Onyx ».



1.b. DMRS de la phrase « Onyx, the Swiss system ».

[1] http://sweaglesw.org/linguistics/ace/

4 Conclusion

Notre méthode prend des éléments partagés sémantiquement comme base pour diviser et reformuler une phrase à partir de sa représentation sémantique DMRS qui donne une annotation combinant sémantique et syntaxe. Nous avons prévu que notre système traite la double négation. Une phrase comme « *It is not impossible* » est considérée comme complexe du fait que le « cumul » de négations pose problème de compréhension des propositions niées pour un faible lecteur par exemple qui serait confus de voir s'il s'agit d'une proposition négative ou positive.

Références

Copestake, A. (2002). *Implementing typed feature structure grammars* (Vol. 110). Stanford: CSLI publications.

Copestake, A. (2009, March). Invited Talk: Slacker semantics: Why superficiality, dependency and avoidance of commitment can be the right way to go. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)* (pp. 1-9).

Guillaume, B., Bonfante, G., Masson, P., Morey, M., & Perrier, G. (2012, June). Grew: un outil de réécriture de graphes pour le TAL (Grew: a Graph Rewriting Tool for NLP)[in French]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 5: Software Demonstrations* (pp. 1-2).

Narayan, S., & Gardent, C. (2014, June). Hybrid simplification using deep semantics and machine translation. In *The 52nd annual meeting of the association for computational linguistics* (pp. 435-445).

Narayan, S., & Gardent, C. (2015). Unsupervised sentence simplification using deep semantics. *arXiv preprint arXiv:1507.08452*.

Saggion, H. (2017). Automatic text simplification. *Synthesis Lectures on Human Language Technologies*, *10*(1), 1-137.

Siddharthan, A. (2014). A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, *165*(2), 259-298.

Woodsend, K., & Lapata, M. (2011, July). Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 409-420).

Zhu, Z., Bernhard, D., & Gurevych, I. (2010, August). A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)* (pp. 1353-1361).

Spécialisation de modèles neuronaux pour la transcription phonémique : premiers pas vers la reconnaissance de mots pour les langues rares

Cécile Macaire[1, 3] Guillaume Wisniewski[2] Séverine Guillaume[1]
Benjamin Galliot[1] Guillaume Jacques[4] Alexis Michaud[1] Solange Rossato[3]
Minh-Châu Nguyên[1, 3] Maxime Fily[1]

(1) Langues et Civilisations à Tradition Orale (LACITO), Unité Mixte de Recherche 7107 CNRS - Sorbonne Nouvelle - Institut National des Langues et Civilisations Orientales (INALCO)
(2) Laboratoire de Linguistique Formelle (LLF), Unité Mixte de Recherche 7110 CNRS - Université de Paris
(3) Laboratoire d'Informatique de Grenoble (LIG), Unité Mixte de Recherche 5217 CNRS - Université Grenoble Alpes - Grenoble INP - Institut national de recherche en informatique et en automatique (INRIA)
(4) Centre de Recherches Linguistiques sur l'Asie Orientale (CRLAO), Unité Mixte de Recherche 8563 CNRS - École des Hautes Études en Sciences Sociales - Institut National des Langues et Civilisations Orientales
cecile.macaire@univ-grenoble-alpes.fr, Guillaume.Wisniewski@u-paris.fr, severine.guillaume@cnrs.fr, b.g01lyon@gmail.com, rgyalrongskad@gmail.com, alexis.michaud@cnrs.fr, Solange.Rossato@univ-grenoble-alpes.fr, minhchau.ntm@gmail.com, maxime.fily@gmail.com

RÉSUMÉ

Nous décrivons les résultats les plus récents que nous avons obtenus dans le cadre du développement d'outils de Traitement Automatique des Langues (TAL) pour réduire l'effort de transcription et d'annotation que doivent fournir les linguistes « de terrain » au fil de leur travail de documentation et description de langues rares. En particulier, nous montrons comment une nouvelle approche neuronale fondée sur la spécialisation d'un modèle de représentation générique permet d'améliorer significativement la qualité de la transcription phonémique automatique, et surtout d'envisager la reconnaissance automatique de mots, approchant ainsi du stade de la *reconnaissance automatique de la parole* au sens plein du terme.

ABSTRACT

A first step towards automatic word recognition for low-resource languages

We describe the latest results we have obtained in the development of NLP (Natural Language Processing) tools to reduce the transcription and annotation workload of field linguists, as part of workflows to document and describe the world's languages. We show how a new deep learning approach based on the fine-tuning of a generic representation model allows to significantly improve the quality of automatic phonemic transcription, and, more significantly, to take a first step towards automatic word recognition for low-resource languages.

MOTS-CLÉS : documentation linguistique assistée par ordinateur, reconnaissance automatique de la parole, modèles neuronaux, science ouverte, linguistique de terrain.

KEYWORDS: Computational Language Documentation, Automatic Speech Recognition, Open Science, Deep Learning, linguistic fieldwork.

1 Introduction

L'amélioration significative de la qualité des outils de Traitement Automatique des Langues (TAL) ouvre de nouvelles perspectives pour faciliter le travail des linguistes de terrain (Anastasopoulos et al., 2020; Partanen et al., 2020; Hjortnaes et al., 2021). Dans nos travaux précédents (Wisniewski et al., 2020a; Michaud et al., 2018), nous avons montré que les méthodes fondées sur les réseaux de neurones permettent de développer des systèmes de reconnaissance phonémique qui aident à la transcription. Ces méthodes sont désormais intégrées dans l'outil ELPIS (Foley et al., 2018), doté d'une interface graphique conviviale (Wisniewski et al., 2020b).

Nous décrivons ici les résultats les plus récents que nous avons obtenus dans le cadre du développement d'outils de TAL pour réduire l'effort d'annotation des linguistes de terrain. Nous montrons comment une nouvelle approche neuronale fondée sur la spécialisation d'un modèle de représentation générique *(fine-tuning)* permet d'améliorer encore la qualité de la transcription phonémique, et surtout de passer à la reconnaissance automatique d'entités de plus haut niveau, à savoir des mots.

2 Spécialisation de modèles pour la transcription phonémique

Principe L'approche mise en œuvre repose sur la spécialisation d'un modèle de représentation multilingue du signal, une méthode introduite par Conneau et al. (2020) pour développer des modèles de reconnaissance de la parole à partir de peu de données.

Dans une première étape, `XLSR-53`, un modèle multilingue appris de manière non supervisée sur un corpus regroupant 56 000 heures d'enregistrements en 53 langues, est utilisé pour construire automatiquement une représentation du signal. Dans une seconde étape, ces représentations sont utilisées en entrée d'un système de reconnaissance phonémique, entraîné à partir de données associées à une transcription manuelle fournie par le/la linguiste.

Utilisation pour la prédiction de phonème Nous avons simplement défini un jeu d'étiquettes correspondant à l'ensemble des caractères composant les phonèmes. Nos expériences passées (Wisniewski et al., 2020a) ont en effet montré que la prédiction des caractères composant les phonèmes (et non pas directement des phonèmes) permettait d'obtenir de bonnes prédictions tout en faisant l'économie de l'étape qui consiste à lister explicitement les phonèmes de la langue. À ce jeu d'étiquettes s'ajoute l'espace, pour délimiter les mots, et par là, s'approcher un peu plus du développement d'un véritable système de reconnaissance de la parole pour les langues rares.

3 Résultats expérimentaux

Nous avons testé la méthode décrite ci-dessus sur deux langues minoritaires de Chine : le na et le japhug. Cela soulève plusieurs défis. Tout d'abord, la quantité de données disponible est très faible. Certes, parmi les langues rares, ces deux langues ne sont pas les moins bien documentées, loin de là. Les corpus transcrits, disponibles dans Zenodo (Galliot et al., 2021) et dans la collection Pangloss (Michaud et al., 2016), sont conséquents : de l'ordre de 3h30 pour le na et 32h pour le japhug. Il faut toutefois mettre ces chiffres en rapport avec les tailles de corpus utilisées pour les langues « courantes » : le corpus libre COMMONVOICE contenait, en 2019, 173h d'audio annoté

pour le français et 780h pour l'anglais (Ardila et al., 2020) ; le modèle de représentation de la parole le plus récent de Facebook, XSL-R, est appris sur 436 000h (quatre cent trente-six mille heures !) d'audio, regroupant 128 langues (Babu et al., 2021), soit une moyenne de plus de 3 000h par langue (moyenne qui cache certes de grandes disparités, mais fournit un ordre de grandeur). En outre, le japhug et le na possèdent des caractéristiques structurelles propres. Par exemple, le système tonal du na (Michaud, 2017) a une organisation fondamentalement différente de celui des 2 langues tonales (sur 53) du corpus multilingue utilisé (`XLSR-53`) : le mandarin et le vietnamien ; et le japhug présente un degré de complexité morphosyntaxique particulièrement impressionnant au vu de son contexte aréal (Jacques, 2021).

La qualité de notre système est évalué en utilisant deux métriques usuelles : le taux d'erreur sur les caractères, *character error rate* (CER), distance d'édition entre la référence et la prédiction calculée au niveau des caractères, et le taux d'erreur sur les mots, *word error rate* (WER), une métrique similaire calculée au niveau des mots.

Le tableau 1 présente les principaux résultats (pour toutes précisions, on se permet de renvoyer à Macaire 2021). Il en ressort que l'approche proposée permet d'obtenir des transcriptions phonémiques de bonne qualité. Le CER pour les deux langues est inférieur à 8%, soit une réduction de 4 points pour le japhug et de 6 points pour le na par rapport aux précédents résultats (Wisniewski et al., 2020b), lesquels reposaient sur une méthode de transcription phonémique qui était également fondée sur un réseau de neurones, mais qui apprenait une représentation du signal uniquement à partir des données d'apprentissage, sans utiliser un modèle pré-entraîné. Il faut toutefois noter que l'erreur au niveau des mots est bien plus élevée que l'erreur au niveau des caractères, mais cette différence est essentiellement liée à la manière dont les deux mesures d'évaluation sont définies : dans la mesure où il y a nettement moins de mots dans une phrase que de caractères, une erreur au niveau d'un caractère (qui se traduit naturellement par une erreur au niveau du mot le contenant) aura un impact beaucoup plus fort sur le WER que sur le CER. Une analyse plus fine des résultats montre que nos systèmes ne font que très peu d'erreurs sur les frontières de mots, aussi bien pour le na que pour le japhug (près de 90% des espaces sont correctement prédits).

	taille corpus apprentissage (mn)	taille corpus test (mots)	WER (%)	CER (%)
na				
évaluation	180	—	41.5	7.9
correction	180	71	38.5	5.7
japhug				
évaluation	600	—	18.5	7.4
correction	600	236	5.4	1.3

TABLE 1 – Résultats obtenus en spécialisant les représentations construites par `XLSR-53` pour la transcription phonémique. Les hypothèses sont évaluées soit par rapport à une référence pré-existante (condition *évaluation*) soit par rapport à une référence obtenue en corrigeant les prédictions du système (colonne *correction*).

Les linguistes de l'équipe ont corrigé quelques transcriptions automatiques. Cette expérience-pilote n'était pas systématisée comme celle de Sperber et al. (2017) (ou d'autres études des processus de *post-édition* en traduction automatique). Elle ne concerne que 71 mots pour le na, et 236 mots pour le japhug (voir, à nouveau, le tableau 1). Elle débouche néanmoins sur une observation claire : le

nombre de corrections à effectuer est beaucoup plus faible que ne le suggère le taux d'erreurs (CER).

Cette observation, bien que peu surprenante (des résultats similaires ont été observés dans le cas de l'évaluation de la traduction automatique), est particulièrement importante : elle suggère que la qualité « réelle » des systèmes est plus élevée que ne le suggéraient les métriques d'évaluation employées jusqu'ici. Au moins dans le cas du na et du japhug, l'effort demandé pour corriger des transcriptions automatiques est considéré comme très faible par les cinquième et sixième auteurs du présent travail (qui sont les linguistes ayant recueilli les corpus na et japhug).

La qualité des prédictions au niveau des mots reste nettement en-deçà de celle obtenue sur des langues « bien dotées ». Ces résultats nous paraissent néanmoins remarquables, et tout à fait encourageants pour l'avenir des efforts conjoints associant TAListes et linguistes de terrain.

Références

Anastasopoulos, A., Cox, C., Neubig, G., and Cruz, H. (2020). Endangered languages meet modern NLP. In *Proceedings of the 28th International Conference on Computational Linguistics : Tutorial Abstracts*, pages 39–45, Barcelona, Spain (Online). International Committee for Computational Linguistics. https://aclanthology.org/2020.coling-tutorials.7/.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. (2020). Common voice : a massively-multilingual speech corpus. *arXiv preprint arXiv :1912.06670*.

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., and Auli, M. (2021). XLS-R : Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv :2111.09296*.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., and Auli, M. (2020). Unsupervised cross-lingual representation learning for speech recognition. *CoRR*, abs/2006.13979. https://arxiv.org/abs/2006.13979.

Foley, B., Arnold, J., Coto-Solano, R., Durantin, G., and Ellison, T. M. (2018). Building speech recognition systems for language documentation : the CoEDL Endangered Language Pipeline and Inference System (ELPIS). In *Proceedings of the 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU), 29-31 August 2018*, pages 200–204, Gurugram, India. ISCA. https://www.isca-speech.org/archive/SLTU_2018/pdfs/Ben.pdf.

Galliot, B., Wisniewski, G., Guillaume, S., Besacier, L., Jacques, G., Michaud, A., Rossato, S., Nguyên, M.-C., and Fily, M. (2021). Deux corpus audio transcrits de langues rares (japhug et na) normalisés en vue d'expériences en traitement du signal. In *Journées scientifiques du Groupement de recherche "Linguistique informatique, formelle et de terrain" (GDR LIFT)*, Grenoble.

Hjortnaes, N., Partanen, N., and Tyers, F. M. (2021). Keyword spotting for audiovisual archival search in uralic languages. In *Proceedings of the Seventh International Workshop on Computational Linguistics of Uralic Languages*, pages 1–7, Syktyvkar, Russia (Online). Association for Computational Linguistics.

Jacques, G. (2021). *A grammar of Japhug*. Number 1 in Comprehensive Grammar Library. Language Science Press, Berlin.

Macaire, C. (2021). Recognizing lexical units in low-resource language contexts with supervised and unsupervised neural networks. Research report, LACITO (UMR 7107). https://hal.archives-ouvertes.fr/hal-03429051.

Michaud, A. (2017). *Tone in Yongning Na : lexical tones and morphotonology*. Number 13 in Studies in Diversity Linguistics. Language Science Press, Berlin. 10.5281/zenodo.439004.

Michaud, A., Adams, O., Cohn, T., Neubig, G., and Guillaume, S. (2018). Integrating automatic transcription into the language documentation workflow : experiments with Na data and the Persephone toolkit. *Language Documentation and Conservation*, 12 :393–429. http://hdl.handle.net/10125/24793.

Michaud, A., Guillaume, S., Jacques, G., Mac, D.-K., Jacobson, M., Pham, T.-H., and Deo, M. (2016). Contribuer au progrès solidaire des recherches et de la documentation : la Collection Pangloss et la Collection AuCo. In *Journées d'Etude de la Parole 2016*, volume 1, pages 155–163. https://halshs.archives-ouvertes.fr/halshs-01341631/.

Partanen, N., Hämäläinen, M., and Klooster, T. (2020). Speech recognition for endangered and extinct samoyedic languages. *CoRR*, abs/2012.05331. https://arxiv.org/abs/2012.05331.

Sperber, M., Neubig, G., Niehues, J., Nakamura, S., and Waibel, A. (2017). Transcribing against time. *Speech Communication*, 93 :20–30.

Wisniewski, G., Guillaume, S., and Michaud, A. (2020a). Phonemic transcription of low-resource languages : To what extent can preprocessing be automated ? In Beermann, D., Besacier, L., Sakti, S., and Soria, C., editors, *Proceedings of the 1st Joint SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) Workshop*, pages 306–315, Marseille, France. European Language Resources Association (ELRA). https://halshs.archives-ouvertes.fr/hal-02513914.

Wisniewski, G., Michaud, A., Galliot, B., Besacier, L., Guillaume, S., Aplonova, K., and Jacques, G. (2020b). Ouvrir aux linguistes « de terrain » un accès à la transcription automatique. In Poibeau, T., Parmentier, Y., and Schang, E., editors, *2èmes journées scientifiques du Groupement de Recherche Linguistique Informatique Formelle et de Terrain (LIFT)*, pages 83–94, Montrouge, France. CNRS. https://hal.archives-ouvertes.fr/hal-03047148.