LIFT-TAL 2022 Actes des journées jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL)

Leonor Becerra, Benoît Favre, Claire Gardent, Yannick Parmentier

LIFT-TAL 2022

Actes des journées jointes des Groupements de Recherche *Linguistique Informatique, Formelle et de Terrain* (LIFT) et *Traitement Automatique des Langues* (TAL)

Leonor Becerra, Benoît Favre, Claire Gardent, Yannick Parmentier (Éds.)

14 au 15 novembre 2022
Marseille, France

Préface

Pour la première fois, les Groupements de Recherche LIFT (Linguistique Informatique, Formelle et de Terrain) et TAL (Traitement Automatique des Langues) organisent des journées scientifiques communes. L'objectif de ces journées est de favoriser les discussions autour des thématiques complémentaires de ces deux GdR. Deux thèmes spécifiques sont mis en lumière à l'occasion de ces journées : "les biais liés aux ressources dans les systèmes d'apprentissage" et "les langues et tâches peu dotées".

Le programme de ces journées est construit autour de 4 conférences invitées, 4 sessions posters permettant aux auteur.e.s de 25 articles de présenter leurs travaux, et 2 panels de discussions. Les contributions à ces journées sont riches et variées, et couvrent différents domaines comme par exemple :

- l'extraction d'information à partir de corpus
- l'assistance à l'annotation manuelle
- la description de langues peu dotées
- le traitement de la langue des signes
- la génération de texte
- des applications de TAL
- les interfaces langagières pour l'accès à l'information

Ces journées sont soutenues par le CNRS et ouvertes à toutes les personnes intéressées, du monde académique ou du secteur privé, aux chercheuses et chercheurs expérimentés ou en début de carrière (étudiantes et étudiants en master ou doctorat).

En vous souhaitant bonne lecture et de nombreux échanges enrichissants tout au long de ces deux jours.

Leonor, Benoît, Claire et Yannick

Comités

Comité d'organisation

Leonor Becerra, LIS / Aix-Marseille Université

Benoit Favre, LIS / Aix-Marseille Université

Claire Gardent, LORIA CNRS

Yannick Parmentier, LORIA / Université de Lorraine

Comité de programme

Gilles Adda - LISN CNRS

Angelique Amelot - LPP CNRS

Pascal Amsili - LaTTiCe / Sorbonne Nouvelle

Leonor Becerra - LIS / Aix Marseille Université

Frederic Béchet - LIS / Aix Marseille Université

Thierry Charnois - LIPN / University of Paris 13

Jean-Pierre Chevallet - LIG / Grenoble Alpes University

Berthold Crysmann - LLF CNRS / Université Paris Cité

Géraldine Damnati - Orange Labs

Marco Dinarelli - LIG CNRS

Solène Evain - LIG / Grenoble Alpes University

Benoit Favre - LIS / Aix-Marseille Université

Olivier Ferret - CEA List

Karën Fort - LORIA / Sorbonne Université

Abdellah Fourtassi - LIS / Aix-Marseille Université

Gaël Guibon - LORIA / Université de Lorraine

Sylvain Kahane - Modyco / Université Paris Ouest Nanterre

Anaïs Lefeuvre-Halftermeyer - LIFO / Université d'Orléans

Damien Lolive - IRISA / Université Rennes 1

Alexis Michaud - LACITO CNRS

Philippe Muller - IRIT / Toulouse University

Alexis Nasr - LIS / Aix-Marseille Université

Tatiana Nikitina - LLACAN CNRS

Magalie Ochs - LIS / Aix-Marseille Université

Yannick Parmentier - LORIA / Université de Lorraine

Thierry Poibeau - LaTTiCe CNRS

Laurent Prévot - LPP / Aix Marseille Université

Carlos Ramisch - LIS / Aix Marseille University

Emmanuel Schang - LLL / Université d'Orléans

Gilles Serasset - LIG / Grenoble Alpes University

Pascale Sébillot - IRISA / INSA de Rennes

Guillaume Wisniewski - LLF / Université Paris Cité

François Yvon - LISN CNRS

Présentations invitées

Antonios Anastasopoulos (George Mason University, Pittsburgh, USA)

Challenges and Considerations in Building Language Technologies for Local Languages

Résumé : The availability of large multilingual pre-trained language models has opened up exciting pathways for developing NLP technologies for languages with scarce resources. In this talk I will summarize some of my group's recent work on the challenges of handling new, unseen languages, as well as general considerations when working with local languages, based on our close collaboration with communities from around the world.

Rachel Bawden (Inria Paris)

Traduction automatique dans des scénarios à faibles ressources : application à des questions linguistiques

Résumé : Malgré les progrès récents en traduction automatique (TA), notamment depuis la généralisation des approches neuronales en traitement automatique des langues, la TA pour les langues peu dotées reste un sujet de recherche difficile. Des techniques telles que l'augmentation des données, le pré-entraînement et l'apprentissage par transfert permettent d'améliorer les performances sur ces langues. Dans cet exposé, je montrerai comment ces techniques peuvent être adaptées à des tâches légèrement différentes qui sont elles aussi caractérisées par un manque de données d'apprentissage. Je présenterai ainsi des travaux récents qui mettent ces techniques au service de deux problématiques linguistiques : la prédiction de cognats et la normalisation du français moderne (français du 17ème siècle).

Karën Fort (LORIA / Sorbonne Université)

Éthique et TAL : ce dont on parle, ce dont on ne parle plus, ce dont on ne parle pas (un état de l'art)

Résumé : Depuis quelques années, l'éthique est devenue un sujet reconnu dans les domaines de l'IA et plus particulièrement dans le traitement automatique des langues (TAL). Cette évolution récente est due à plusieurs facteurs, dont le fait que le TAL est devenu suffisamment rentable commercialement pour sortir des laboratoires de recherche et envahir nos vies quotidiennes, avec des conséquences immédiatement visibles pour le grand public. Je reviendrai dans cette présentation sur l'évolution qu'a connu le sujet sur la dernière décennie, qui a vu certaines problématiques devenir évidentes (comme la rémunération des travailleurs du clic) et ne plus être discutées, alors que d'autres (notamment les biais des modèles de langues) occupent le devant de la scène, occultant les questions les plus difficiles. Une large place sera laissée à la discussion, afin de permettre des échanges de vues sur ces sujets.

Alexis Michaud (LACITO CNRS)

Documentation des langues et linguistique informatique : présentation de travaux en cours

Résumé : Les méthodes informatiques présentent un fort potentiel pour aider à faire face à l'urgence que constitue la documentation des langues en danger. Les modèles créés par l'apprentissage machine peuvent aider à accomplir des tâches chronophages telles que la transcription et la traduction mot à mot (réalisation de gloses interlinéaires). Mais le traitement automatique des langues reste peu utilisé dans la documentation linguistique, notamment parce que la technologie est encore nouvelle (et en évolution rapide), et que la prise en main des outils demande des compétences spécialisées. L'exposé reviendra sur des travaux en cours dans trois directions, deux applicatives et la troisième plus fondamentale. La première consiste à créer une interface conviviale à l'usage des non-informaticiens. La seconde : mettre en place des équipes pluridisciplinaires qui puissent déployer, à l'échelle, les meilleurs outils, tirant parti du libre accès aux données favorisé par la transition en cours vers les pratiques de Science ouverte. La troisième direction explorée consiste à tirer parti des modèles statistiques au profit de la recherche fondamentale (en linguistique aussi bien qu'en informatique), en sondant les représentations neuronales.

Table des matières

Production et Perception de ⟨’⟩ chez les Tsimane’

William N. Havard[1, 2] Camila Scaff[3, 1] Loann Peurey[1] Alejandrina Cristia[1]
(1) Language Acquisition Across Cultures Team, Laboratoire de Sciences Cognitives et de Psycholinguistique, Département d’Etudes cognitives, ENS, EHESS, CNRS, PSL University, France
(2) Cognitive Machine Learning Team, INRIA, Paris, France
(3) Humam Ecology Group, Institute of Evolutionary Medicine, University of Zurich, Switzerland
william.havard@gmail.com, camila.scaff@cri-paris.org,
loannpeurey@gmail.com, alecristia@gmail.com

RÉSUMÉ

Le tsimane’ est une langue parlée en Bolivie par plusieurs milliers de personnes et pourtant la phonologie du tsimane’ n’a pas été décrite en détail. Avec ce projet, nous souhaitons faire un pas vers une meilleure description en nous concentrant sur un aspect de la langue qui nous semble particulièrement inhabituel : le son réprésenté orthographiquement par ⟨’⟩, traditionellement analysé comme un coup de glotte /ʔ/. Nous avons émis l’hypothèse que, en termes techniques, ⟨’⟩ est réalisé comme un son glottal battu. Nous avons enregistré deux locuteurs adultes de Tsimane’ produisant des paires (quasi) minimales impliquant ce son. Dans cet article, nous présentons des analyses centrées sur une syllabe extraite de six paires minimales : /ki-kiʔ/. Les analyses des spectrogrammes ont suggéré qu’un locuteur utilisait systématiquement la glottalisation de la voyelle et, dans une moindre mesure, une occlusion totale, alors qu’elles étaient ambiguës chez notre autre informateur. Cependant, la présentation des syllabes cibles à ces deux informateurs ainsi qu’à deux autres locuteurs adultes de Tsimane’ a révélé des preuves claires qu’ils pouvaient clairement retrouver la syllabe voulue. Ensemble, ces données suffisent à écarter notre hypothèse initiale d’un flap glottal, puisqu’une fermeture n’était jamais évidente chez l’un des locuteurs, et suggèrent plutôt qu’un ensemble plus complexe d’indices acoustiques peut être à la disposition des auditeurs.

ABSTRACT

A study of the production and perception of ⟨’⟩ in Tsimane’.

Tsimane’ is a language spoken in Bolivia by several thousand people and yet the phonology of Tsimane’ has not been described in detail. With this project, we want to take a step towards better description by focusing on an aspect of language that we find particularly unusual : the sound represented in spelling with ⟨’⟩, usually analyzed as a glottal stop /ʔ/. We hypothesized that ⟨’⟩ is a glottal flap. We recorded two adult speakers of Tsimane’ producing (near-)minimal pairs involving this sound. In this paper, we present analyses focused on a syllable extracted from six minimal pairs : /ki-kiʔ/. Analyses of the spectrograms suggested one speaker consistently used vowel glottalization and to a lesser extent closure, whereas these were ambiguous in our other informant. However, presentation of the key syllables to these two informants and two other adult Tsimane’ listeners revealed clear evidence that they could clearly recover the intended syllable. Together, these data suffice to rule out our initial hypothesis of a glottal flap, since a closure was never obvious in one of the speakers, and suggests instead a more complex set of acoustic cues may be at listeners’ disposal.

MOTS-CLÉS : phonologie, perception, production, adaptation des expériences de laboratoire.

KEYWORDS: phonology, perception, production, adapted lab experiments.

Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain *(LIFT) et* Traitement Automatique des Langues *(TAL),*
Marseille, France, 14 au 15 novembre 2022 , pages 1–8,

1 Introduction

There are few linguistic descriptions of Tsimane' (Sakel, 2011; Gill, 1999; Gill & Gill, 1999; Ritchie, 2017a,b), and none covers the phonology of Tsimane' very well. In this paper, we describe a small-scale research project aimed at studying the sound marked in the orthography as ⟨'⟩.

⟨'⟩ has been described to us as a glottal stop (Ritchie, undocumented personal communication, 2018), which has been well documented across languages. However, both distributional and phonetic evidence do not point to this really being a glottal stop. In distributional terms, glottal stops are hard to hear, so in most languages, they occur only at word and syllable onsets. When a language allows it in syllable offset, it also allows it in syllable onset – again suggesting that glottal stops will only appear in salient positions. In contrast, in Tsimane', ⟨'⟩ occurs *only* in syllable offsets. Moreover, it doesn't seem to be a "real consonant" distributionally, because Tsimane' allows very few consonants in syllable offset, and never allows two consonants together in that position. And yet, ⟨'⟩ can follow the nasal consonants /m, n/ or approximants /r, ʊ/ at the end of syllables, suggesting that it doesn't "count" as a consonant from a syllable structure point of view.

Regarding phonetic evidence, based on preliminary discussions with Tsimane' research assistants and others, ⟨'⟩ does not sound like full glottalization of the preceding vowel or nasal (i.e., it does not sound like the vowel or nasal has a different vocal quality caused by the glottis vibrating in a specific way during the sound) but instead, that it is a "glottal flap/tap" (where the air is stopped at the glottis more briefly than for a stop, or perhaps not at all).

Based on these descriptions and considerations, we designed a small-scale study to be carried out during a 6-week visit to San Borja, a city close to traditional Tsimane' territory. Our original hypothesis was that ⟨'⟩ is phonologically a feature of the preceding sound (i.e., the vowel or nasal sound /m, n, r, ʊ/ that immediately precedes ⟨'⟩) and phonetically a glottal flap cued primarily by a change in the vowel or nasal formants. This hypothesis needed to be tested via acoustic analyses of speech samples and perceptual reports by informants.

Although there exist some speech corpora in Tsimane', none of them have the high audio quality required for our analyses nor for eliciting judgments by native Tsimane' informants. Therefore, our first aim during our short visit was to collect good quality audio recordings of a small number of informants producing the target sound ⟨'⟩ in a given context, as well as the same context without that sound. The second aim was to present extracts from those recordings, potentially edited to neutralize certain acoustic cues, to the same or other native Tsimane' informants, in order to assess the extent to which they rely on that acoustic information in their perception of ⟨'⟩ being present or not. Given time constraints, our perceptual studies do not include edited stimuli, but only natural stimuli. We return to this below.

2 Production study

2.1 Methods

2.1.1 Equipment and procedure

For the recordings, informants were seated in the same room as the experimenter and were provided with Tsimane' phrases to read, each containing one of the target words from our set of minimal pairs. We intended to use LIG-Aikuma (Gauthier *et al.*, 2016) to collect this data, using a simple smartphone. However, due to low audio quality (16kHz) as well as random recording failures, we implemented a computer-based version that re-implements the text-elicitation and respeaking modes and allows to record at a 44.1kHz sampling-rate. [1]

Participants recorded sentences in a quiet room using a Dell Precision 3561 Computer running Ubuntu 20.04.5 LTS. Participants use JBL Quantum 300 headphones, equipped with a foam windscreen to record the stimuli. Headsets were connected to the computer via a USB audio cable adapter.

2.1.2 Stimuli

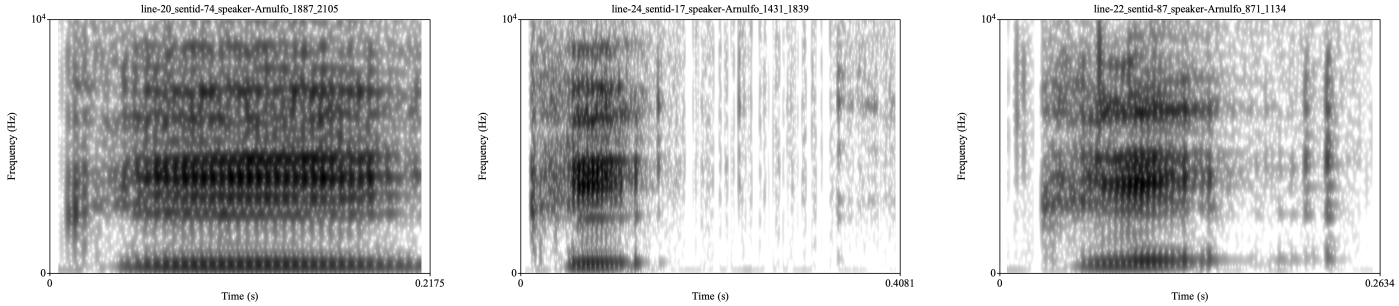
The words were chosen to elucidate the phonetic properties of ⟨'⟩ in a minimally varying context, so we intended to find minimal pairs. Our analyses of the Tsimane' dictionary initially revealed a certain number of meaningful minimal pairs. However, in consultation with our two informants, many of these items were found to have been mistranscribed in the dictionary, or to be variants of each other. Thus, in collaboration with our two informants, we identified 14 minimal pairs, and 6 near-minimal pairs (in which one more sound differed between the paired items). Additional work allowed us to identify more (near-)minimal pairs, and to construct meaningless minimal pairs. However, analyses here will focus on the 6 meaningful minimal pairs first identified where ⟨'⟩ occurred in the syllable "qui" /ki/, so as to control for context and lexical status.

Each of the two items of each pair was recorded in five contexts : isolation, a natural phrase (drawn from the dictionary or grammar, or elicited from our informants) and three carrier sentences, where the position of the target word varies are used, so as to mitigate the effects of prosodic variations (e.g. "TARGET-WORD mo' nash peyacdye' yu yi", which means "TARGET-WORD is the word I am saying" ; "yu ra' yi TARGET-WORD jeñej peyacdye'", which means "I will say TARGET-WORD as a word" ; "yu ra' yi mo' peyacdye' TARGET-WORD ", which means "I will say the word TARGET-WORD ").

2.1.3 Informants

This work was possible thanks to an extensive collaboration with two informants, both of whom are native Tsimane' speakers. They also use Spanish frequently, and this was the language in which all studies were conducted. They are both literate, and very experienced in collaborating with researchers. One of them had previously worked as an informant with a linguist ; the other has over 10 years of experience as a translator and research assistant with a team of biological anthropologists. They grew up in two different villages, and our interactions with them in the transcription and analysis of Tsimane' materials strongly suggested to us that there are idiosyncratic or dialectal differences in

1. Code available at https://github.com/William-N-Havard/williaikuma



(a) /ki/ (b) /ki/ (c) /ki?/

FIGURE 1 – A salient contrast between /ki/ (left) and /ki?/ (right), as well as an ambiguous case (/ki/ intended). The example on the left has no glottalization in the vowel or apparent glottal closure/release. The example on the right has both clear glottalization in the vowel and closure followed by release. The example in the middle seems to have some glottalization towards the end of the vowel, and what could seem like a very long closure and a release.

their production and potentially their perception. As we explain below, this was confirmed for the contrast of interest here.

This study was approved by CER U-Paris 2022-84-CRISTIA. Informants were compensated for their participation in this work, as well as in other ongoing work, at a rate of 145 Bolivianos per day.

2.1.4 Analyses

We call a "clip" each audio recording of one item in one context (isolation, natural sentence, or carrier phrase). First, we used Praat to inspect spectrograms of over 100 items (including ⟨'⟩ or not, in a variety of vocalic and sonorant contexts) by our first informant, which made it obvious that checking for formant transitions visually would be very challenging. Instead, we focused on two aspects : (a) the presence of glottalization in the vowel, as defined perceptually as well as in the spectrogram (compare Figures 1a and 1c) ; and (b) the presence of a clear closure followed by a release (see Figure 1c). Third, we segmented 1. the syllable, starting before the release of the /k/ and ending with the vowel /i/ or, if present, a potential closure followed by release ; 2. the vowel /i/ (from the first to the last voiced period) ; 3. if present, the closure followed by release if present. A Praat script was used to extract the duration of the vowel and, if an unambiguous closure was present, its duration.

2.2 Results

We annotated 136 clips from our two informants (68 clips each). Table 1 shows the prevalence of vowels judged to be glottalized, and closures judged to be present, for each speaker separately. From this analysis, it appears that one of our two informants consistently uses vowel glottalization (unambiguously present in 83% of /ki?/ clips, and unambiguously absent in 84% of /ki/ clips) ; and he often also signals a closure, although closures were ambiguous present in half of /ki/ clips, meaning that the speaker paused after /ki/. We found vowel glottalization status to be ambiguous in nearly half of the clips corresponding to our second informant. However, when glottalization was not ambiguous, it was clearly absent for /ki/, and more commonly present than absent for /ki?/. This second informant never produced unambiguous closures, regardless of the target.

	Glottalization			Closure		
Target	Present	Absent	Ambiguous	Present	Absent	Ambiguous
/ki/	1 / 0	31 / 22	6 / 16	0 / 0	20 / 23	18 / 15
/ki?/	25 / 13	0 / 3	5 / 14	21 / 0	2 / 13	7 / 17

TABLE 1 – Perceptually-determined presence of vowel glottalization and closure (followed by release), as a function of target. Each cell indicates the prevalence separately for each of our two informants.

We then analyzed vowel duration separately for our two informants. In the first informant, an unpaired Welch two-sample t-test revealed a significant difference in vowel duration across /ki/ and /ki?/ : $t(64.3) = 4.3$, $p < .001$; 119 ms in /ki/ versus 80ms in /ki?/. In the second informant, the difference in vowel duration was not significant : $t(63.4) = 1.0$, $p = .32$; 171 ms in /ki/ versus 163ms in /ki?/. Finally, although we have nothing it to compare this to, we thought it may still be informative to report on the duration of the closure (plus release) in our first speaker (since the second did not produce unambiguous closures) : The mean closure (and release) duration was 179 ms (range 64-298ms).

2.3 Discussion

Due to logistical constraints, we could only collect production data from two adult Tsimane' speakers, who frequently use Spanish. Nonetheless, they are both native and fluent Tsimane' speakers, highly experienced in research, and detailed-oriented, leading us to be confident that their production reflects clear speech. In these data, it appears to be the case that the contrast is stronger in one of our speakers than the other. Whether this is due to idiosyncratic or dialectal variation needs to be established in further work by interviewing more speakers. In the speaker where the contrast was clearly present, it was obvious in vowel glottalization, but also in the frequent presence of a closure.

3 Perception study

Given the natural variation in cue presence across our two speakers, we believe our perception study, based only on natural syllables (without artificial cue manipulation) can be quite informative. So as not to confuse listeners, we separated tokens by speaker, and collected perceptual reports through two separate listening studies.

3.1 Methods

3.1.1 Equipment and procedure

Praat was used to present the stimuli and collect judgments. Participants were tested one at a time in a quiet room. Participants sat in front of the computer and wore JBL Quantum 300 headsets at full power. There were three experiments presented in succession. The first was simply to acquaint participants with the procedure : two maximally different syllables (/ko⟨'⟩/ and /yi/) were extracted from the speech of our first informant. These two clips were presented a total of three times in a random order. On the screen, participants could see the syllables "co'" and "yi", as well as a button that

allowed them to listen to the same stimulus up to three times before making a decision. We explained the procedure and stayed with them while they did this. We then answered any questions that they could have, and provided them feedback on the procedure if needed. The next two experiments were self-paced. In one, they heard all the stimuli for our first informant, and they heard all the stimuli for the second informant.

3.1.2 Stimuli

We extracted all /ki/ and /kiʔ/ syllables from the phrases analyzed above, except for two syllables of our second speaker that contained a click. Due to an error, some syllables were extracted several times with slightly different segmentations, resulting in 138 stimuli to be presented, 72 from our first speaker and 66 from the second.

3.1.3 Participants

Our two informants participated, as well as two other native Tsimane' speakers, who worked as research assistants for another project. Three of them were male, one female. All are young to middle-aged adults (one born in the 70', one in the 80's, and the last two, which includes the female participant, were born in the 90's). All come from different communities. This study was approved by the CER U-Paris 2022-84-CRISTIA. Participants were compensated 145 Bolivianos as part of their daily workday. The perception study as a whole took about 15 minutes.

3.2 Results

Table 2 shows the percent of correct responses (i.e., listener reported hearing a /kiʔ/ when the target was /kiʔ/ and vice versa) separated by participant, speaker, and whether the speaker intended /ki/ or /kiʔ/. It is obvious that all for listeners were overwhelmingly capable of retrieving the intended syllable. This is remarkable given that there was wide variation in implementation, as described above.

	Speaker 1		Speaker 2	
Participant	/ki/	/kiʔ/	/ki/	/kiʔ/
1 (m)	85	100	95	93
2 (m)	95	97	100	83
3 (m)	82	72	86	72
4 (f)	60	100	70	69

TABLE 2 – Percent of correct responses, separated by participant (rows), stimulus speaker and target (columns). Participants 1 and 2 correspond to our first and second informants of the Production study described above.

3.3 Discussion

Despite variation in the input, participants were well above chance, suggesting they probably used a mix of cues, some beyond vowel glottalization and closure. We know that in other contrasts, a multitude of acoustic cues can correspond to several alternate gestures. Here, listeners could be relying to differences in formant structure, formant transitions, and even pitch levels.

4 General discussion

Our hypothesis was that ⟨'⟩ was actually glottalization in the vowel or sonorant. Our production experiment revealed this was not necessarily the case, since one of the speakers didn't unambiguously glottalize or use modal voicing for his vowels. Moreover, although the other speaker did, he also produced clear closures and releases in a majority of his items. These data also allowed us to rule out clearly the hypothesis whereby the standard pronunciation for the contrast relied on the presence of a glottal flap or stop, since one of our two speakers never systematically and unambiguously use full closures.

The perceptual study showed that both speakers were signaling the contrast, since all four listeners (two who had participated in the production study, and two who had not collaborated with us on this project) were well above chance level for both speakers and both items of the pair. In the future, analyses using random forests on a variety of acoustic cues could help us understand the acoustic indices listeners were employing for each one of our speakers. In addition, analyses could inspect whether relevant acoustic cues vary as a function of the position in which the target syllable occurred. We noticed very strong reduction at the end of phrases, particularly for our first informant, and furthermore in this position closures are likely less salient. Our stimuli allowed the word containing the target syllable to occur sentence-initially and sentence-medially before a sonorant (vowel or semi-vowel), which should allow us to further study potential effects on the following sonorant.

The present study has several limitations. We were only able to study production in two native speakers, who also have extensive experience with Spanish, and it is unclear to what extent their use of this second language may affect their pronunciation. That said, there is no particular reason to suppose that Spanish could affect it, since glottal stops and vowel glottalization do not occur in any salient way in the local variety of Spanish. Another limitation was the use of elicited phrases, and the focus on one specific syllable, /ki/, as context. Future studies can employ the corpus of production we have collected to generalize this procedure to other contexts. We also hope to utilize other extant corpora to study this contrast in a wider variety of materials. Our perceptual study is also affected by a relatively small sample size, although results are quite clear in that they suggest the contrast is easily recovered by listeners. Nonetheless, sample size will be a limitation when attempting to use these data to infer which acoustic cues are most useful to listeners. Our data also suggest there is some idiosyncratic and/or dialectal variation, for which greater sample sizes and purposeful sampling of different geographic areas would be ideal. Finally, taking into account the goal of establishing what is the phonological category underlying this contrast, instrumental studies may be necessary, so as to recover the precise gestures that are made, although transporting such equipment to field conditions may be challenging.

Remerciements

We thank our informers, Arnulfo Cary and Manuel Roca, without whom this study would have been impossible. Funding for this study comes from the J. S. McDonnell Foundation (Understanding Human Cognition Scholar Award) ; European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (ExELang, Grant agreement No. 101001095).

Références

GAUTHIER E., BLACHON D., BESACIER L., KOUARATA G.-N., ADDA-DECKER M., RIALLAND A., ADDA G. & BACHMAN G. (2016). LIG-AIKUMA : A mobile app to collect parallel speech for under-resourced language studies. In *Interspeech 2016 (short demo paper)*.

GILL W. (1999). A pedagogical grammar of the Chimane (Tsimane') language. *San Borja, Bolivia : New Tribes Mission*.

GILL W. & GILL R. (1999). Chimane-English Dictionary. *San Borja, Bolivia : New Tribes Mission*.

RITCHIE S. (2017a). Agreement with the internal possessor in chimane* : A mediated locality approach. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, **41**(3), 660–716. DOI : https ://doi.org/10.1075/sl.41.3.05rit.

RITCHIE S. (2017b). Posesión y Relactiones Gramaticales en Chimane. *Lenguas Indígenas de Bolivia : Teoría y Práctica*, p.7̃.

SAKEL J. (2011). *A grammar of Mosetén*, volume 33. Walter de Gruyter.

Abstraction ou hallucination ? État des lieux et évaluation du risque pour les modèles de génération de résumés automatiques de type séquence-à-séquence

Eunice Akani
Aix-Marseille Univ, CNRS, LIS, Marseille, France
Enedis, Marseille, France
eunice.akani@lis-lab.fr

RÉSUMÉ

Nous présentons un résumé étendu de l'article (Akani *et al.*, 2022) présenté à la conférence TALN 2022.

ABSTRACT

Abstraction or Hallucination ? Status and Risk assessment for sequence-to-sequence Automatic Text Summarization Models..

We present an extended abstract of the paper (Akani *et al.*, 2022) that was presented at the 2022 TALN conference.

MOTS-CLÉS : Résumé automatique de texte, hallucination, mesure d'évaluation.

KEYWORDS: Automatic text summarization, hallucination, evaluation metric.

1 Résumé étendu

Le résumé automatique de texte par abstraction consiste à générer automatiquement une synthèse d'un document en capturant ses informations importantes. Cette tâche a connu un regain d'intérêt grâce à l'arrivée des modèles Transformers et des modèles de langue pré-entraînés (Vaswani *et al.*, 2017; Devlin *et al.*, 2019). Malgré ces avancées, la tâche reste difficile surtout quand il s'agit d'un résumé par abstraction — résumé un texte en utilisant de nouveaux mots ou des paraphrases. Cao *et al.* (2018) ont montré que 30% des résumés produits par une sélection de systèmes de résumé automatique par abstraction contiennent des informations incohérentes vis-à-vis du document source qui ne sont pas capturées par les métriques d'évaluation habituelles telle que ROUGE (Lin, 2004). Maynez *et al.* (2020) ont qualifié ces incohérences d'hallucinations. Ils définissent une hallucination comme étant une information se trouvant dans le résumé que l'on ne peut déduire à partir du document source. Sur cette base, deux types d'hallucinations sont évoquées : les hallucinations *intrinsèques*, définies comme des mots ou groupes de mots du résumé qui sont tirées du document mais qui ne sont pas déductibles, et les hallucinations *extrinsèques*, qui introduisent des informations hors document. Alors que les hallucinations intrinsèques peuvent générer des contre-sens et être considérées comme des erreurs, les hallucinations extrinsèques sont souvent des abstractions ou introductions d'informations dont on ne peut directement vérifier la factualité qu'à partir d'une connaissance générale dépassant le document à résumer. Une illustration de ces phénomènes est donnée dans l'exemple de la table 1. En analysant

DOCUMENT : Il tourna sept films de la saga, dont "L'espion qui m'aimait", "Rien que pour vos yeux" et "Octopussy". Outre sa carrière cinématographique, Roger Moore s'était illustré au début de sa carrière dans plusieurs séries télévisées comme "Ivanhoé", "Le Saint" ou "Amicalement vôtre", où il partageait l'affiche avec Tony Curtis. Avec plus de cinquante films à son actif, Roger Moore avait quelque peu délaissé le grand écran ces dernières années. Ses dernières apparitions se sont faites essentiellement dans des téléfilms ou des séries. En 2003, il est fait chevalier commandeur de l'Ordre de l'Empire britannique et obtient également, en 2008, le titre de Commandeur des Arts et des Lettres décerné par la France. Très sensible à la cause animale, il soutenait activement l'association PETA. Après trois divorces, Roger Moore était marié depuis 2002 à une richissime danoise, Kristina Tholstrup.	**REF :** Roger Moore [s'est éteint]. [L'acteur britannique] connu pour [son élégance en toutes circonstances et son humour] avait endossé [le costume de James Bond] [de 1973 à 1985].
	PTGEN : "roger moore, association peta, [est décédé] [dimanche] [à l'âge de 85 ans], [a annoncé sa famille à l' afp] .l'association peta a également [fait part de son passage] à roger moore, "" ivanhoé""."
	C2C : [L'acteur britannique] Roger Moore, connu notamment pour son rôle dans "L'espion qui m'aimait", [est décédé] [mardi] [à l'âge de 87 ans], [a annoncé sa famille à la télévision].
	BARTHEZ : L'acteur et [réalisateur américain] Roger Moore, [décédé] [à l'âge de 95 ans], est connu pour son rôle dans "L'espiosn qui m'aimait", "Rien que pour vos yeux" et "Octopussy".
	MT5 : Le [cinéaste] [britannique] Roger Moore [est décédé] [mardi] [à l'âge de 77 ans], [a annoncé son avocat] Tony Curtis.

TABLE 1 – Exemple de sorties des différents systèmes (PTGEN (See *et al.*, 2017), C2C (Martin *et al.*, 2020), BARTHEZ (Kamal Eddine *et al.*, 2021) et MT5 (Xue *et al.*, 2021)) utilisés sur un document du corpus « Orange-Sum Abstract » (Kamal Eddine *et al.*, 2021). La référence est annotée suivant la typologie des abstractions tandis que les résumés candidats sont annotés suivant la typologie des erreurs. Le code couleur est le suivant : AbsNInf, HorsDoc, NonInf et le gris pour les entités hors du document. Le résumé généré par PTGEN est grammaticalement incorrect.

L'acteur britannique Roger Moore, connu notamment pour son rôle dans "L'espion qui m'aimait", est décédé mardi à l'âge de 87 ans a annoncé sa famille à la télévision.

FIGURE 1 – Exemple des types d'hallucination définit par Maynez *et al.* (2020). Résumé d'un article de « Orange-Sum Abstract » généré par le système C2C. En bleu les hallucinations intrinsèques : informations tirées du document mais non inférables à partir de celui-ci. Et en rouge les hallucinations extrinsèques : informations n'étant pas mentionnées dans le document.

plus en détail le système C2C (Figure 1), on remarque que le résumé produit mentionne que « l'acteur est britannique », ce qui n'est pas précisé dans le document. En effet, il est bien mentionné qu'il a été fait chevalier commandeur de l'Ordre de l'Empire britannique mais cela ne signifie pas qu'il est britannique ; c'est donc une hallucination intrinsèque car plusieurs éléments du document sont associés sans être pour autant déductibles à partir de celui-ci. Aussi, le système précise que « l'acteur est décédé mardi à l'âge de 87 ans ». N'étant pas dans le document c'est donc une hallucination extrinsèque. Pour tenter de limiter la production d'hallucinations, Maynez *et al.* (2020) ont utilisé le « textual entailment », c'est-à-dire la capacité d'un texte à inférer un autre texte, comme mesure de sélection d'un résumé le plus aligné à la source possible. Ces travaux ont montré que le « textual entailment » n'était pas une mesure suffisante pour garantir la fidélité du résumé par rapport au document source. Durmus *et al.* (2020) ont proposé d'exploiter un système de question-réponse pour évaluer la factualité d'un résumé, en générant des questions à partir des phrases du résumé et en vérifiant que leur réponse dans le document est identique à celle à l'origine de la question. Bien que plusieurs études aient été publiées sur les hallucinations dans le cadre de résumés automatiques en anglais, il n'y en a aucune sur le français.

L'étude porte sur l'analyse de sorties de systèmes de résumé de l'état de l'art sur un corpus français, Orange-Sum (Kamal Eddine *et al.*, 2021), afin de savoir à quel point les modèles sous-jacents sont sujets aux hallucinations. Ainsi, nous avons d'abord introduit une typologie des erreurs et abstractions contenues dans les résumés. Puis, nous nous sommes concentrons sur les hallucinations extrinsèques, et en particulier sur les entités nommées qui peuvent apparaître dans un résumé hypothèse mais pas dans le document source. Ceci nous a permis de comparer les productions de systèmes de génération RNN [1] et transformers [2] à l'aide d'une analyse manuelle de leurs erreurs, et de proposer une mesure d'évaluation du risque potentiel d'erreurs (le *risque d'hallucination*) lorsqu'un modèle essaie de faire des abstractions sur les entités.

L'analyse des sorties des différents systèmes a montré que malgré des scores ROUGE très intéressants, ces systèmes sont encore affectés par de nombreuses erreurs liées à la factualité des informations qu'ils présentent. Notre mesure d'évaluation du risque d'hallucination sur les entités nommées a permis de constater que les systèmes séquence à séquence à base de transformers prennent énormément de risque en essayant de prédire des entités hors du document. Aussi, l'analyse du corpus Orange-Sum montre les limites de ce corpus. En effet, ces résumés de référence contiennent des informations non présentes dans les documents.

Les typologies ainsi que les détails des résultats obtenus sont disponibles dans l'article original Akani *et al.* (2022).

Références

AKANI E., FAVRE B. & BECHET F. (2022). Abstraction ou hallucination ? état des lieux et évaluation du risque pour les modèles de génération de résumés automatiques de type séquence-à-séquence. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale / Travaux originaux*, p. 1–10, Avignon, France : Association pour le Traitement Automatique des Langues. Abstraction or Hallucination ? Status and Risk assessment for sequence-to-sequence Automatic.

CAO Z., WEI F., LI W. & LI S. (2018). Faithful to the original : Fact aware neural abstractive summarization. *ArXiv*, **abs/1711.04434**.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DURMUS E., HE H. & DIAB M. (2020). FEQA : A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 5055–5070, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.454.

KAMAL EDDINE M., TIXIER A. & VAZIRGIANNIS M. (2021). BARThez : a skilled pretrained French sequence-to-sequence model. In *Proceedings of the 2021 Conference on Empirical Methods*

1. Modèle à base de pointeur-générateur (See *et al.*, 2017)
2. CamemBERT2CamemBERT (Martin *et al.*, 2020; Scialom *et al.*, 2020), Barthez (Kamal Eddine *et al.*, 2021) et mT5 (Xue *et al.*, 2021)

in Natural Language Processing, p. 9369–9390, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.740.

LIN C.-Y. (2004). ROUGE : A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, p. 74–81, Barcelona, Spain : Association for Computational Linguistics.

MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2020). Camembert : a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

MAYNEZ J., NARAYAN S., BOHNET B. & MCDONALD R. (2020). On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 1906–1919, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.173.

SCIALOM T., DRAY P.-A., LAMPRIER S., PIWOWARSKI B. & STAIANO J. (2020). Mlsum : The multilingual summarization corpus. *arXiv preprint arXiv :2004.14900*.

SEE A., LIU P. & MANNING C. (2017). Get to the point : Summarization with pointer-generator networks. In *Association for Computational Linguistics*.

VASWANI A., SHAZEER N. M., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention is all you need. *ArXiv*, **abs/1706.03762**.

XUE L., CONSTANT N., ROBERTS A., KALE M., AL-RFOU R., SIDDHANT A., BARUA A. & RAFFEL C. (2021). mT5 : A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 483–498, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.41.

Alignement des embeddings des définitions et du contexte pour un assistant de lecture sensible au contexte

Ioana Ivan Nathan Chometton
Aix-Marseille Université, CNRS, LIS, Marseille, France
ioana.ivan@etu.univ-amu.fr, chometton.n@gmail.com

RÉSUMÉ

La désambiguïsation lexicale (word sense disambiguation ou WSD) est une tâche du traitement automatique de la langue qui vise à identifier, à partir d'un mot, de son contexte d'occurrence et d'une liste de sens possibles, le sens du mot le plus adapté. Cette tâche pourrait permettre le développement de liseuses plus sophistiquées que les liseuses actuelles en présentant, lorsqu'on clique sur un mot pour en connaître le sens, le sens le plus probable étant donné le contexte. Nous allons dans cette étude proposer quelques pistes pour aborder ce problème.

ABSTRACT

Matching contextual and definitional embeddings for a sense-aware reading assistant.

WSD is a branch of natural language processing which aims to identify, by means of the context of occurence and a list of possible senses, the most suitable sense of the word. This task could enable the development of more sophisticated e-readers which, while clicking on a word to find out its meaning, would present the reader with the most probable meaning given the context. In this study, we will propose some pointers to tackle this problem.

MOTS-CLÉS : Wiktionnaire, désambiguïsation lexicale, word2vec, fastText, flauBERT.

KEYWORDS: Wiktionary, WSD, word2vec, fastText, flauBERT.

1 Introduction

La désambiguïsation lexicale est une branche du domaine du traitement automatique de la langue qui vise a identifier, à partir du contexte d'apparition d'un mot et d'une liste des sens possibles, le sens du mot le plus adapté. Dans cet étude nous utilisons des méthodes sans apprentissage, fondées sur l'hypothèse qu'il existe une similarité entre le contexte d'apparition d'un mot et sa définition. L'implémentation de cette idée remonte à l'algorithme introduit par Michael Lesk en 1986, dont une version simplifiée consistait à compter le nombre de mots en commun entre le contexte et la définition d'un mot. Mais elle constitue aussi une source d'inspiration pour des approches modernes en WSD, comme GlossBERT (Huang *et al.*, 2019), un modèle neuronal qui apprend si une définition correspond ou pas à un contexte donné.

La tâche de désambiguïsation suppose que l'on dispose d'un corpus dans lequel les mots sont associés à leur sens en contexte. De tels corpus sont peu courants car très chers à développer. Une manière de contourner le problème consiste à utiliser un dictionnaire donnant pour chaque entrée une définition et pour chaque définition un ou plusieurs exemples. Ces exemples comportent une occurrence de l'entrée

Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain *(LIFT)* et Traitement Automatique des Langues *(TAL)*,
Marseille, France, 14 au 15 novembre 2022 , pages 13–19,

qui est désambiguïsée (son sens est la définition sous laquelle l'exemple apparaît). L'ensemble des exemples constitue alors un corpus où chaque exemple comporte un mot désambiguïsé. Nous avons suivi Segonne *et al.* (2019) et avons utilisé Wiktionary. D'après les auteurs, cette ressource, malgré son caractère dynamique et collaboratif, permet d'obtenir de meilleures performances, parmi les différentes ressources existant en français, pour la tâche de désambiguïsation.

2 Le jeu de données

Notre jeu de données est obtenu à partir de l'extrait ontolex Dbnary du 29-05-2022 du Wiktionnaire français (Sérasset, 2015) [1]. Le fichier Dbnary a été traité pour obtenir un jeu de données organisé en 6 colonnes : lemme, définition, exemple, index du lemme dans l'exemple, partie du discours du lemme, registre(s). Dans ce jeu de données, chaque ligne correspond à un exemple, comme illustré dans le Tableau 1.

Mot	Définition	Exemple	Index	POS	Registres
pile	Précisément	Il est minuit pile	3	adv	Familier|Populaire

TABLE 1 – Exemple d'une entrée du jeu de données.

Les exemples correspondant aux mots monosémiques (qui ont un seul sens/définition dans le Wiktionary) ont été enlevés du corpus car ce cas de figure ne nous intéresse pas. En effet, au vu d'une application concrète d'aide à la lecture sur liseuse, ces mots ne posent aucun problème et leur traitement est trivial. Nous avons également retiré les mots outils ainsi que les lemmes associés à des définitions ne contenant pas d'exemples. Le jeu de données final se compose donc de **34 026 lemmes**, **102 138 définitions** et **188 000 exemples**.

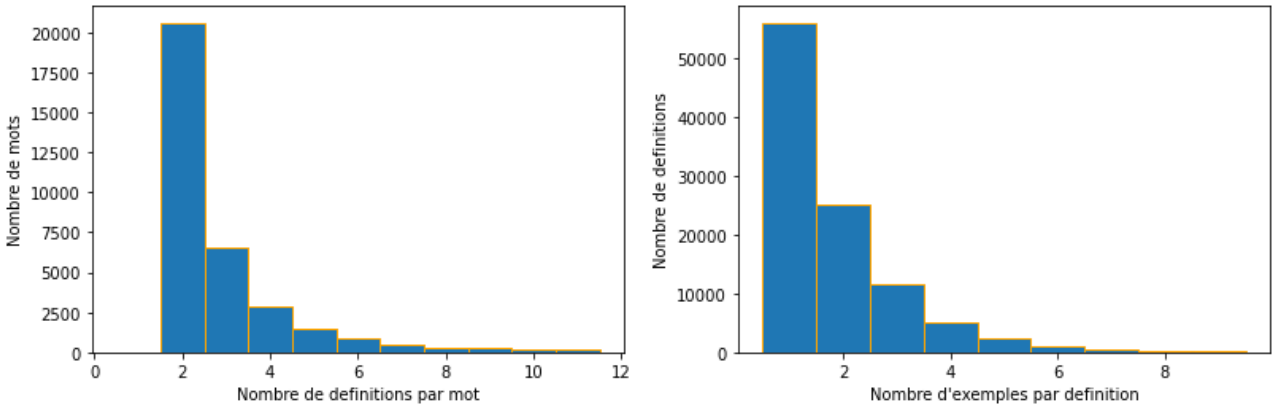


FIGURE 1 – **A gauche** : Histogramme de polysémie (nombre de définitions par lemme), **A droite** : Histogramme du nombre d'exemples par définition

Le nombre moyen de définitions par lemme est de **3** (Figure 1). Remarquons, pour commencer, que la plupart des lemmes ont 2 définitions, même si un certain nombre (764) compte plus de 10 définitions (allant jusqu'à 60 définitions différentes pour un même lemme). En ce qui concerne le nombre d'exemples par définition la moyenne est de **1,84**.

1. http://kaiko.getalp.org/about-dbnary/download/

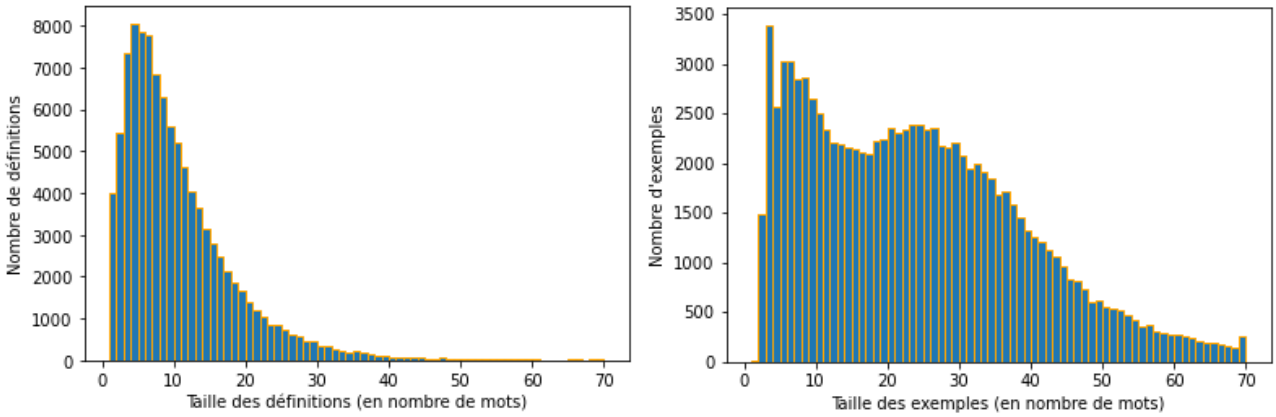


FIGURE 2 – **A gauche** : Histogramme du nombre de mots par définition, **A droite** : Histogramme du nombre de mots par exemple.

La longueur des exemples est assez élevée, avec une moyenne de **24 mots par exemple** (Figure 2). Ceci est probablement dû aux citations assez nombreuses dans notre base de données. La taille des définitions est plus courte avec une moyenne de **10 mots par définition**. On remarque qu'on compte très peu de définitions longues (15 mots ou plus), mais on compte plus de **30 000 définitions courtes** (5 mots ou moins).

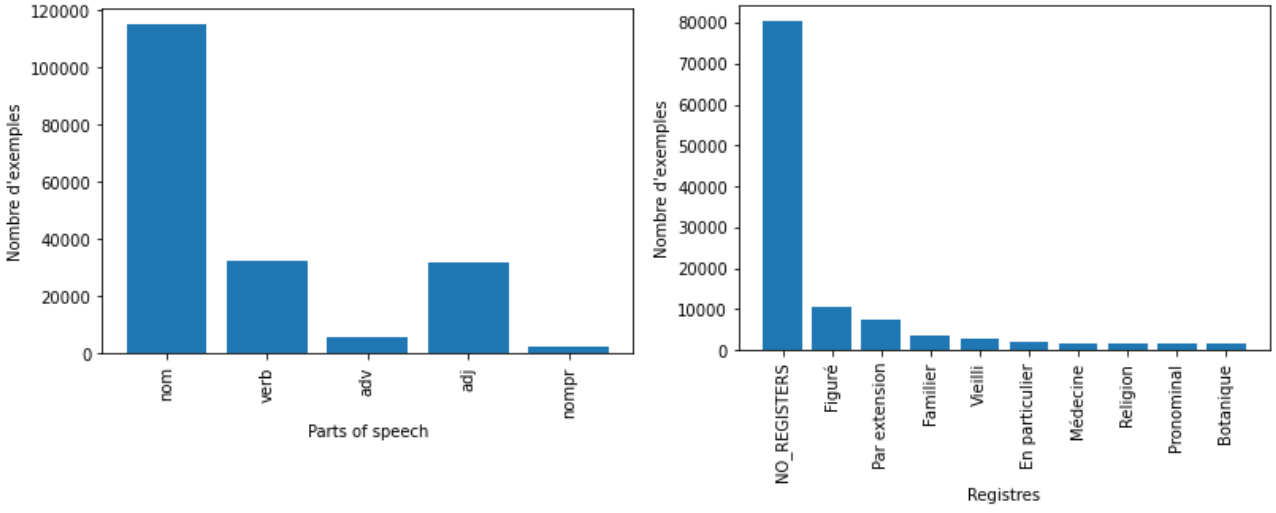


FIGURE 3 – **A gauche** : Histogramme des parties de discours, **A droite** : Histogramme des registres

Concernant les parties de discours, on constate une prédominance assez attendue des noms qui correspondent à **60%** (115 079) des exemples du corpus (Figure 3). Les adjectifs et les verbes représentent quant à eux autour de **30%** (64 726) des exemples. En ce qui concerne les registres, on observe que 42% des exemples ont au moins un registre associé. Les plus fréquents sont : "figuré", "par extension", "familier" et "vieilli".

3 Expériences et résultats

Pour réaliser la désambiguïsation, nous avons utilisé des méthodes sans apprentissage, fondées sur la similarité cosinus entre une représentation vectorielle du contexte d'occurrence d'un mot et la représentation vectorielle de sa définition. Notre modèle *baseline* consiste simplement à additionner les embeddings des mots de la définition d'une part et ceux du contexte d'autre part, et comparer les deux embeddings résultants.

Nous avons réalisé les expériences sur un jeu de données test constitué de 18 800 exemples, soit 10% du jeu de données. Les exemples du jeu de test sont prises aléatoirement dans le jeu de données complet. Nous gardons 90% des données pour pouvoir appliquer des méthodes fondées sur l'apprentissage à l'avenir.

Nous avons commencé en utilisant des embeddings **fastText** de taille 300 fournis par Grave *et al.* (2018). En appliquant notre modèle baseline, nous avons obtenu un score de **35.69%**. Étant donné la faiblesse de ce résultats, nous l'avons comparé au choix aléatoire. Pour déterminer le score du choix aléatoire, nous avons choisi au hasard pour chaque exemple une des définitions du mot-cible et nous avons calculé l'exactitude. Nous avons obtenu une exactitude de **32.59 %**.

Considérant que notre baseline améliorait très faiblement les résultats du choix aléatoire, nous avons tenté de l'améliorer en donnant plus de poids à certains mots dans la phrase. Dans ce but, nous avons conçu deux modèles, basés sur les hypothèses suivantes : i) dans une définition, les mots qui aident le plus à différencier le sens sont les premiers mots de la phrase (il s'agit souvent d'hyperonymes) et ii) dans un exemple, ce sont les mots positionnés le plus proche du mot-cible qui aident le plus à distinguer le sens.

Les détails des deux modèles sont donnés ci-dessous :

1. **Modèle début de la définition.** Le poids du mot dépend de la longueur de la phrase et décroît linéairement avec la position dans la phrase, pondérée par un facteur α.

$$w_1(i) = n - i\alpha,$$

   où $n \in \mathbb{N}^*$ est le nombre de mots de la phrase, $\alpha \in (0, 1]$ est une constante fixée et $i \in [0, n)$ est la position du mot dans la phrase.

2. **Modèle contexte dans l'exemple.** Les voisins du mot cible ont un poids égal à la taille de la phrase et le poids décroît linéairement vers les deux extrémités de la phrase.

$$w_2(i) = n + 1 - \alpha * |i_c - i|,$$

   où $n \in \mathbb{N}^*$ est le nombre de mots de la phrase, $\alpha \in [0, 1)$ est une constante fixée, i_c est la position du mot cible et $i \in [0, n)$ est l'index du mot

Modèle	Exactitude
Aléatoire	32.59%
Baseline	**35.69%**
Début déf. ($\alpha = 0.1$)	35.65%
Contexte ex. ($\alpha = 0.3$)	**35.79%**

TABLE 2 – Exactitude des modèles avec des embeddings fastText

En étudiant le tableau 2 qui présente les résultats, nous remarquons une amélioration très faible (0.1%) lors de la pondération, mais globalement les scores sont très proches du modèle baseline. L'exactitude ne semble pas s'améliorer significativement en changeant les poids en accord avec ces hypothèses. En effet, ni les exemples ni les définitions ne semblent pas suivre une des hypothèses que nous avons formulé.

Au vu de ces résultats, nous avons décidé de garder le même modèle baseline et de changer de type d'embedding. Nous avons choisi deux autres types d'embeddings : des embeddings non contextuels provenant de word2vec, un précurseur de fastText, et des embeddings contextuels provenant de FlauBERT.

Dans le cas de word2vec, nous avons considéré plusieurs modèles pré-entraînés sur le corpus frWac2Vec fournis par Fauconnier (2015). Nous avons choisi celui qui maximise l'écart entre la similarité cosinus entre les exemples avec leurs définitions associées d'un coté, et la similarité des exemples avec une définition du mot tirée au hasard d'un autre coté. Le meilleur modèle de ce point de vue a été le modèle cbow avec des vecteurs de taille 200.

Dans le cas de FlauBERT, nous avons utilisé les embeddings de taille 786 fournis par Le *et al.* (2020). Pour les définitions, ce sont les embeddings CLS qui ont été utilisés, et pour les exemples, les embeddings contextuels générés pour les tokens dans la position du mot cible.

Une comparaison des résultats du modèle baseline avec les trois types d'embeddings est montrée dans le tableau 3. Les embeddings word2vec ont la meilleure exactitude et dépassent considérablement leur compétiteurs. Les résultats de FlauBERT sont surprenants, car ils se situent bien en deçà de word2vec et sont très proches du choix aléatoire. Cela pourrait indiquer que les embeddings CLS et les embeddings contextuels ne sont pas proches du point de vue de la similarité cosinus et qu'une autre façon d'utiliser les embeddings FlauBERT obtiendrait des résultats plus satisfaisants.

Modèle	Exactitude
Aléatoire	32.59%
FlauBERT	33.06%
fastText	35.69%
word2vec	**48.24%**

TABLE 3 – Exactitude du baseline avec des embeddings différents

4 Analyse

Pour mieux comprendre nos résultats, nous avons regardé la performance du meilleur modèle : celui basé sur les embeddings word2vec, sur les parties de discours et les registres des mots. Nous avons remarqué que les noms propres ont la meilleure exactitude - **62%**, alors que les adjectifs et les verbes sont les plus difficiles à identifier, avec une exactitude autour de **44%** (Tableau 4).

De manière similaire, nous avons regardé la performance du modèle par rapport au registre (tableau 5) : les mots sans registre ont la meilleure exactitude - **45%**. Parmi les registres les plus représentés, le registre *Par extension* semble être un des plus difficiles à identifier.

Embedding	Noms	Adjectifs	Verbes	Adverbes	Noms propres
Aléatoire	32.38 %	35.27 %	29.12 %	34.50 %	47.77 %
word2vec	50.22 %	44.03 %	44.88 %	45.22 %	**61.94 %**

TABLE 4 – Exactitude part of speech

Embedding	Pas de registre	Figuré	Par extension	Familier
Aléatoire	32.33 %	29.32%	26.78%	34.29%
word2vec	45.07 %	45.17%	41.88%	43.51%

TABLE 5 – Exactitude registres

5 Conclusions

Les pistes d'amélioration et d'analyse sont nombreuses. On aimerait, entre autres, déterminer si le corpus a une granularité adéquate en terme de nombre de définitions par mot et comprendre pourquoi les embeddings flauBERT n'obtiennent pas de bons résultats. Étendre nos méthodes à des réseaux de neurones en bénéficiant de la supervision du Wiktionary est aussi une des pistes envisagées.

Remerciements

Nous remercions nos professeurs, Alexis Nasr et Carlos Ramisch pour leur aide et leur bienveillance.

Références

FAUCONNIER J.-P. (2015). French word embeddings.

GRAVE E., BOJANOWSKI P., GUPTA P., JOULIN A. & MIKOLOV T. (2018). Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

HUANG L., SUN C., QIU X. & HUANG X. (2019). GlossBERT : BERT for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3509–3514, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1355.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020). Flaubert : Unsupervised language model pre-training for french. In *Proceedings of The 12th Language Resources and Evaluation Conference*, p. 2479–2490, Marseille, France : European Language Resources Association.

SEGONNE V., CANDITO M. & CRABBÉ B. (2019). Using Wiktionary as a resource for WSD : the case of French verbs. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, p. 259–270, Gothenburg, Sweden : Association for Computational Linguistics. DOI : 10.18653/v1/W19-0422.

SÉRASSET G. (2015). DBnary : Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. *Semantic Web – Interoperability, Usability, Applicability*, **6**(4), 355–361. DOI : 10.3233/SW-140147, HAL : hal-00953638.

Apprentissage actif pour l'extraction des aspects explicites : application à des avis non annotés en français

Maroua Bouadabous[1] Anna Pappa[1]
(1) LIASD, Université Paris 8, 2 rue de la liberté, 93526 Saint Denis, France
m.boudabous@ai.univ-paris8.fr, ap@up8.edu

RÉSUMÉ

Dans cet article, nous présentons un processus de bout en bout qui utilise l'apprentissage actif pour améliorer l'extraction des aspects explicites pour des langues à faibles ressources dans le cadre d'analyse des opinions et des sentiments. L'extraction des aspects explicites ou implicites, reste difficile en raison de la rareté des données labellisées et de la complexité du processus d'annotation manuelle. Nous avons conçu un processus en deux étapes : nous commençons avec une pré-labellisation via un CRF en utilisant l'apprentissage par transfert de connaissances. Ensuite, nous utilisons l'apprentissage actif pour améliorer la performance de labellisation, gérer les labels manquants et réduire les efforts d'annotation humaine. Nous avons utilisé deux corpus, composés des avis utilisateurs en langue française, sur des produits de beauté et des appareils électroniques pour l'évaluation des processus. Les résultats montrent que l'étape d'apprentissage actif améliore les performances de labellisation lorsque 30% des labels initiaux sont corrigés.

ABSTRACT

Active Learning for Explicit Aspect Term Extraction : a use case for unlabeled French reviews

In this paper, we present an end-to-end process using active learning to improve explicit aspect extraction for low-resource language in opinion and sentiment analysis. Handling Aspect Term Extraction (ATE) as a supervised sequence labeling task remains challenging due to labeled data scarcity and the complexity of the manual annotation process. We define a two-step process starting from defining a pseudo-labeler CRF using cross-domain learning. Then, we use active learning to deal with label uncertainty resulting from the first step and to reduce human annotation efforts. Two web-scraped datasets of French reviews on beauty products and electronic devices were used for process evaluation. Results show that the active learning step improves the learning model's performance when 30% of initial labels are corrected.

MOTS-CLÉS : Apprentissage actif, Extraction d'aspect explicite, Apprentissage inter domaines.

KEYWORDS: Active learning, Explicit Aspect Term Extraction, Cross-domain learning.

1 Introduction

Le World Wide Web est un gisement de données numériques couvrant une multitude de domaines d'intérêt (économie, politique, etc.) d'une richesse et d'une facilité d'accès sans précédent. En effet, l'explosion de l'usage d'internet engendre des volumes croissants de données langagières et multilingue disponibles en ligne. Ces données proviennent des sources différentes comme les réseaux sociaux, les plateformes de services ou les sites de commerces en ligne. Durant les dernières années,

Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain *(LIFT) et* Traitement Automatique des Langues *(TAL),*
Marseille, France, 14 au 15 novembre 2022 , pages 20–28,

les entreprises se sont intéressées à les exploiter afin d'y extraire des connaissances pertinentes pour l'amélioration de leurs produits et/ou services. Cet intérêt s'est également manifesté dans la communauté scientifique du traitement automatique des langues (TAL). L'apprentissage automatique et statistique est un des outils parmi les plus utilisés, pour analyser et traiter les données textuelles préalablement annotées. Cependant, les données étiquetées (ou labellisées) sont inégalement disponibles, en fonction de la langue. Nous désignons par langue à faible ressources, les langues ayant peu ou pas de données labellisées disponibles.

Dans cet article, nous nous intéressons à l'identification des aspects permettant la reconnaissance des termes utilisés pour désigner les attributs et caractéristiques des produits ou des services à partir des avis des utilisateurs. Cette tâche n'est pas triviale, elle est coûteuse, chronophage et suscite une attention considérable pour la création des modèles performants sur des données non labellisées. Nous distinguons deux types d'aspects : explicites et implicites. Dans ce qui suit, nous considérons l'identification des aspects explicites à partir de données textuelles en Français comme exemple de langue à faibles ressources. Ce contexte nous a motivé à proposer un processus d'identification des aspects adapté aux langues à faibles ressources. Notre processus de traitement se définit en deux étapes d'apprentissage : la première consiste à la pré-labellisation des données par transfert de connaissances et la deuxième à l'amélioration de cette labellisation par apprentissage actif.

Nous apportons des modifications au niveau des deux étapes du processus pour s'adapter à la rareté des données labellisées. En effet, nous facilitons le transfert de connaissances, durant la phase de pré-labellisation, par la suppression des descripteurs dépendants du domaine d'application notamment les noms (e.g "restaurant" , "cuisine", "nourriture", "plats") et les verbes (e.g "déguster","savourer"). Nous agissons également sur la stratégie de sélection pour l'apprentissage actif, par la définition de l'indice de confiance. Ceci est la moyenne pondérée des probabilités de l'exactitude des termes identifiés comme "aspect" qui composent un avis utilisateur. Cette méthode de calcul permet de gérer le déséquilibre de fréquence entre termes "aspects " et "non-aspects".

La description de la méthode est organisée comme suit. La section 2 introduit la tâche d'identification des aspects dans la littérature. La section 3 décrit les différentes étapes du processus proposé pour la reconnaissance des aspects à partir de corpus non labellisés. Dans la section 4, nous présentons les données utilisées pour l'évaluation du processus , les expérimentations et les résultats. La section 5 conclut l'article et propose quelques perspectives.

2 État de l'art

L'identification des aspects présente une étape clé dans l'analyse des sentiments et des tendances. Or il s'agit d'une tâche complexe notamment pour les langues qui possèdent peu ou pas de données labellisées. L'utilisation d'apprentissage statistique supervisé a suscité de l'intérêt pour l'identification des aspects en définissant cette tâche sous forme d'un problème de labellisation séquentielle. Différentes architectures de réseaux profonds ont été utilisées pour résoudre le problème d'identification des aspects à savoir les réseaux récurrents de type LSTM (Liu *et al.*, 2015), le mécanisme d'attention (Li *et al.*, 2018), les réseaux de convolution CNN (Xu *et al.*, 2018) ainsi que les transformateurs (Santos *et al.*, 2021). Les performances de toutes ces architectures dépendent de la disponibilité des données d'apprentissage labellisées, ce qui rend leur utilisation problématique pour les langues à faibles ressources.

La rareté des données labellisées pour entraîner les modèles d'apprentissage supervisés a été considérée par (Li *et al.*, 2020) qui ont eu recours à l'augmentation des données. Ils génèrent des données supplémentaires à travers un modèle de génération masqué conditionnel appliqué sur les données d'apprentissage. Cette méthode est assez incontrôlable car elle risque de changer la sémantique des opinions générés. (Chen & Qian, 2020) ont utilisé des prototypes logiciels appris sur des données externes générées par des modèles de langage pré-entraînés pour pallier ce problème. Cette approche nécessite cependant des jeux de données volumineux pour entraîner les modèles de langages. Les auteurs soulignent que plus le volume des données est important mieux est la performance de leur modèle. Dans cet article, nous considérons la situation extrême où aucune donnée labellisée dans le domaine d'application n'est disponible. Le processus que nous proposons repose sur une étape de pré-labellisation qui permettra ensuite d'entraîner le modèle d'apprentissage via l'apprentissage actif pour améliorer l'identification des aspects par le biais de plusieurs itérations de correction des labels et d'enrichissement des données.

L'apprentissage actif (AA) représente une alternative qui permet de réduire l'effort de labellisation manuelle au profit de l'amélioration des performances.

Dans le domaine de traitement automatique des langages, AA a été considéré pour différentes tâches de labellisation séquentielle à savoir la reconnaissance d'entité nommée (NER) (Shelmanov *et al.*, 2019) et l'étiquetage morpho-syntaxique (PoS) (Ringger *et al.*, 2007). Dans ce contexte, l'AA a été dans un premier temps utilisé avec des modèles d'apprentissage automatique standards (Settles & Craven, 2008) (Settles, 2009). Ensuite, l'AA a été associé à des modèles d'apprentissage profonds (Schröder & Niekler, 2020) ; (Ren *et al.*, 2021), (Shen *et al.*, 2018). Cette association a révélé des résultats intéressants en termes de coût de calcul et de labellisation.

L'apprentissage actif est un processus itératif qui sélectionne un échantillon des données non labellisées selon une stratégie de sélection prédéfinie et invite un expert à les labelliser. Dans cet article, nous adaptons ce processus pour permettre à la fois d'enrichir des données et de corriger les pré-labels. En effet, l'expert est invité à vérifier l'exactitude de la labellisation des données sélectionnées à partir de l'ensemble de données pré-labellisées initial.

Définir la stratégie de sélection est incontournable pour tout processus d'apprentissage actif. La stratégie de sélection par incertitude est l'approche la plus répandue dans le cadre de l'apprentissage actif (Scheffer *et al.*, 2001) ; (Culotta & McCallum, 2005). Elle consiste à sélectionner les instances pour lesquelles le modèle d'apprentissage est le moins confiant. Notre approche redéfinit l'indice de confiance pour la labellisation séquentielle. Pour chaque instance, nous utilisons la moyenne pondérée des indices de confiance de chacun de ses termes "aspects" et "non aspects". La pondération est adoptée pour accorder plus d'importance aux aspects qui sont moins fréquents dans les avis des utilisateurs.

3 Description du processus d'identification des aspects

En effet, nous abordons le problème de la rareté des données labellisées pour l'identification des aspects explicites à partir des avis utilisateurs. Nous considérons le français comme un exemple de langues à faibles ressources. Tout d'abord, les données sont pré-labellisées via le transfert de connaissances par adaptation de domaine à l'aide d'un modèle de champs aléatoires conditionnels CRF. Ensuite, nous entraînons un modèle d'apprentissage profond de type BiLSTM-CNN-CRF pour

la reconnaissance des aspects sur les données précédemment pré-labellisées dans un cadre supervisé. Enfin nous appliquons l'apprentissage actif pour gérer les aspects manquants et incorrectement identifiés et améliorer les performances de pré-labellisation. La figure 1 illustre les différentes étapes du processus proposé.

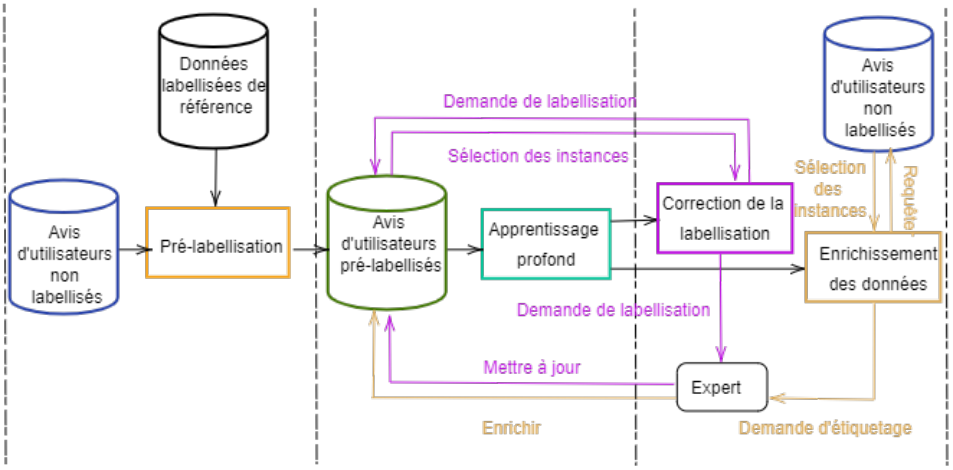


FIGURE 1 – Schéma global du processus proposé pour l'identification des aspects explicites .

Nous proposons la phase de pré-labellisation en tant que étape de pré-traitement pour l'annotation des données non labellisées en utilisant les champs aléatoires conditionnels (CRF). CRF est un modèle de prédiction probabiliste utilisé pour segmenter et labelliser les données séquentielles notamment les textes. Il prédit le label de chaque terme en considérant son voisinage.

Concrètement, les données en entrée pour le CRF sont des vecteurs de caractéristiques qui représentent différentes informations à propos du terme à savoir sa valeur littérale, le nombre de lettres, s'il s'agit d'un terme en majuscule ou minuscule, s'il s'agit d'un signe de ponctuation ou d'un chiffre. L'étiquette morpho-syntaxique (PoS) est également incluse. Nous considérons également les mêmes informations pour les K-mots voisins.

Pour renforcer l'indépendance de notre modèle CRF par rapport au domaine d'application, nous avons omis la caractéristique "valeur littérale" des mots dont l'étiquette morpho-syntaxique correspond aux catégories noms ("Noun") et verbes ("Verb") puisqu'il s'agit de catégories les plus dépendantes du domaine.

L'architecture BiLSTM-CNN-CRF se compose de deux couches de plongements de mots : la première couche utilise l'algorithme Glove, pré-entraîné sur le corpus wikipedia en Français, pour fournir les plongements des mots composant la séquence. La deuxième couche utilise un réseau de neurones convolutifs (CNN) pour extraire les informations morphologiques au niveau du caractère et les encoder dans des représentations neuronales. La sortie de ces couches passe par la suite dans une couche LSTM bidirectionnelle pour capturer des informations sur les mots passés et futurs de la séquence. Enfin, une couche CRF est ajoutée pour produire la séquence la plus probable des labels prédits (Voir la figure 2).

Nous adaptons le processus de l'apprentissage actif, pour permettre à la fois d'enrichir des données et de corriger les pré-labels. En effet, l'expert est invité à vérifier l'exactitude de la labellisation des données sélectionnées à partir de l'ensemble initial de données pré-labellisées.

Nous adaptons la stratégie de sélection pour permettre la classification multi-label et pallier au

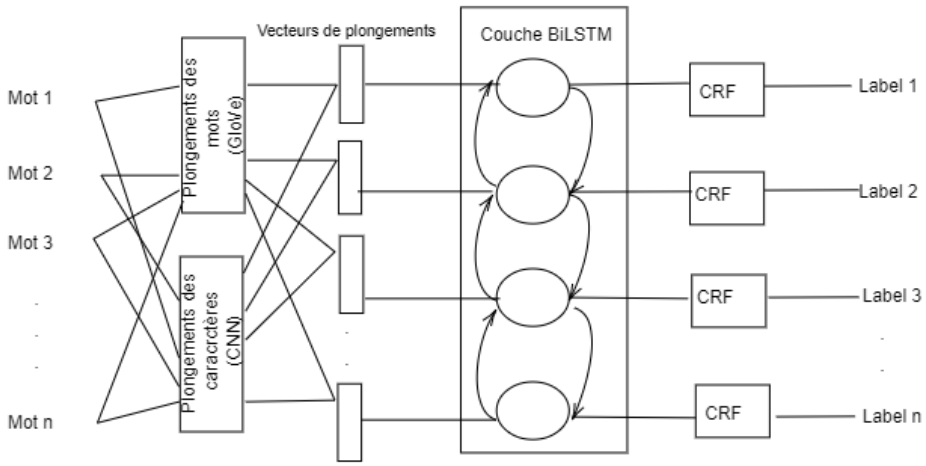


FIGURE 2 – Architecture du modèle d'apprentissage BiLSTM-CNN-CRF

déséquilibre de représentativité des classes «aspect» et «non-aspect». En effet, nous sélectionnons des instances dont la prédiction est la moins confiante. Pour s'adapter au contexte de la labellisation séquentielle, nous associons à chaque instance la moyenne des scores d'incertitude des termes qui la composent. Par ailleurs, les termes qui ne représentent pas des aspects sont supprimés du calcul. Ceci permet d'accélérer la convergence de l'apprentissage actif.

	Paramètre	Valeur
Modèle CRF	Coefficient de Régularisation L1	0.5
	Coefficient de Régularisation L2	0.1
	Nombre maximal d'itérations	50
	Algorithme d'optimisation	lbfgs
Modèle BiLSTM-CNN-CRF	Initialisation des plongements	U(-0.5,0.5)
	Nombre des couches convolutions	3
	Taille des couches convolutions	30
	Dimension de la couche BiLSTM	200
	Taux d'abandon (Dropout)	0.5
	Momentum	0.9
	Taux d'apprentissage	0.01
	Taille du batch	32

TABLE 1 – Définition des paramètres d'apprentissage pour le modèle CRF et le modèle BiLSTM-CNN-CRF

4 Évaluation

Pour évaluer les différentes étapes du processus proposé, nous nous sommes appuyés essentiellement sur deux types de jeux de données : des jeux de données labellisés des avis sur des restaurants (en français), mis à disposition dans la compétition SemEval (SemEval-2016 Tâche 5 «Aspect-Based Sentiment Analysis»), et des corpus non labellisés extraits à partir du web dont deux sous-ensembles ont été manuellement labellisés pour les utiliser comme jeu de données de référence.

Nous évaluons les deux étapes du processus en termes de F1-score qui combine les mesures de précision et de rappel et s'adapte mieux à notre tâche d'identification des aspects où les termes qui ne désignent pas des aspects sont plus présents que ceux les désignant. Pour la pré-labellisation, nous avons réglé les paramètres du CRF en l'entraînant sur les seules données disponibles en français pour l'identification des aspects explicites (Jeu de données sur les restaurants de SemEval 2016). Les valeurs des paramètres retenues sont décrites dans la Table 1.

La figure 3 montre la comparaison entre l'approche utilisant tous les descripteurs de CRF (Version 1, en bleu) à celle favorisant l'adaptation du domaine en omettant le descripteur "valeur littérale" aux termes ayant comme étiquette morpho-syntaxique "Noun" et "Verb" (version 2, en orange). Les résultats confirment que la deuxième approche favorise le transfert des connaissances. Nous remarquons une amélioration de $13,3\%$, $17,5\%$, et $12,5\%$ en termes de F1-score respectivement sur le corpus des musées, celui des appareils technologiques et le corpus des produits de beauté.

A l'issue de la phase de pré-labellisation, nous constatons un taux de labels incorrects qui s'élève à 40% et un taux de labels "aspects" manquants de l'ordre de 59.5% sur le corpus (sans aucune annotation préalable) des produits de beauté (pour rappel le transfert s'effectue à partir des labels du corpus restaurants).

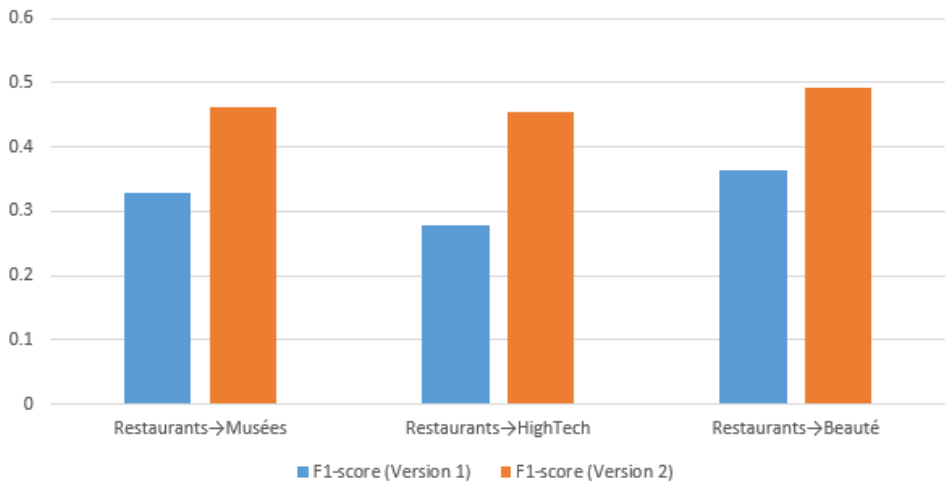


FIGURE 3 – Évaluation de la pré-labellisation par adaptation de domaines en prenant en compte tous les descripteurs (Version 1) et en ajustant les descripteurs (Version 2)

Pour la deuxième étape, nous utilisons le modèle BiLSTM-CNN-CRF comme modèle d'apprentissage profond pour l'identification des aspects explicites par apprentissage actif. Nous initialisons les plongements de mots en utilisant les plongements pré-entraînés de type Glove avec 300 dimensions (GloVe.840B.300d) fournit par (Pennington *et al.*, 2014). Nous définissons également les différents paramètres d'apprentissages comme décrit dans la table 1.

Nous réalisons les expérimentations pour l'apprentissage actif sur deux corpus pré-labellisés composés chacun de 5000 avis utilisateurs qui portent respectivement sur les appareils électroniques et les produits de beauté. Le modèle d'apprentissage BiLSTM-CNN-CRF est initialement entraînés sur 20 époch, puis nous itérons le processus 10 fois.

La figure 4 illustre les résultats d'évaluation, en termes de F1-score, de différentes expérimentations décrites ci-dessus sur les jeux de données des avis concernant les produits de beauté et les produits technologiques. Nous comparons la stratégie de sélection par incertitude que nous avons adaptée à la stratégie de sélection aléatoire. Nous désignons par TCL le taux de correction de labels et par TED

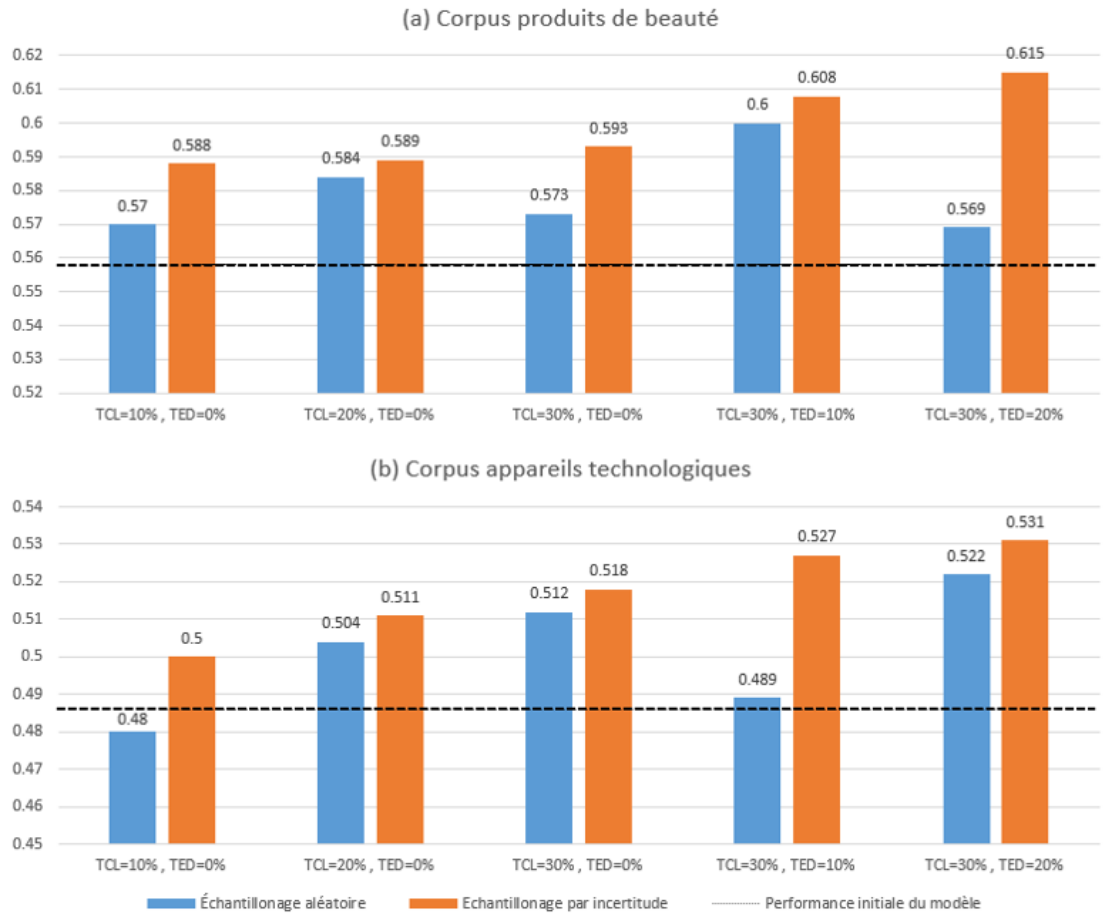


FIGURE 4 – Évaluation de l'apprentissage actif en terme de F1-score sous différentes configurations sur les corpus : (a) produits de beauté et (b) appareils technologiques.

le taux d'enrichissement des données. Nous limitons les valeurs de TCL et le TED à 30% et 20% respectivement, pour garantir un gain de 50% en termes d'effort et de temps nécessaires pour une labellisation totale du jeu de données d'apprentissage. Nous avons également favorisé la correction des labels afin de réduire l'impact des erreurs de pré-labellisation sur la performance du modèle. La ligne en pointillé représente la performance du modèle BiLSTM-CNN-CRF avant l'application de l'apprentissage actif.

Globalement, les résultats soulignent l'intérêt de la mise en place d'un processus d'apprentissage actif puisque nous observons une amélioration autour de 5% et de 6% en terme de F1-score respectivement sur le corpus des avis utilisateurs sur les appareils technologiques et celui sur les produits de beauté par rapport aux résultats obtenus par le modèle BiLSTM-CNN-CRF sans application de l'apprentissage actif (sans enrichir les données ni les corriger par l'expert).

5 Conclusion

Dans cet article, nous avons proposé un processus de bout en bout qui s'appuie sur l'apprentissage actif pour améliorer la labellisation des aspects explicites, pour les langues à faibles ressources. Pour effectuer les tests, nous avons utilisé un corpus composé d'avis utilisateurs (en français) sur les produits de beauté et un autre sur les appareils électroniques, les deux corpus créés à partir de

Web. Notre approche consiste en un processus en deux étapes, commençant par une pré-labellisation pour remédier à la rareté des données labellisées via le transfert de connaissances par adaptation de domaines. Ensuite, le processus d'apprentissage actif est utilisé pour corriger les erreurs de pré-labellisation et réduire ainsi les coûts d'annotation manuelle. Les expériences montrent que l'apprentissage actif améliore considérablement les performances du modèle d'apprentissage lorsque 30% des labels initiaux sont corrigés. Dans les travaux futurs nous adapterons le processus proposé pour permettre l'identification et labellisation des termes d'aspect sur des ensembles de données composés de textes exprimés en différentes langues. De plus, nous ajouterons une autre phase de traitement pour catégoriser les aspects identifiés.

Références

CHEN Z. & QIAN T. (2020). Enhancing aspect term extraction with soft prototypes. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 2107–2117.

CULOTTA A. & MCCALLUM A. (2005). Reducing labeling effort for structured prediction tasks. In *AAAI*, volume 5, p. 746–751.

LI K., CHEN C., QUAN X., LING Q. & SONG Y. (2020). Conditional augmentation for aspect term extraction via masked sequence-to-sequence generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online : Association for Computational Linguistics.

LI X., BING L., LI P., LAM W. & YANG Z. (2018). Aspect term extraction with history attention and selective transformation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*.

LIU P., JOTY S. & MENG H. (2015). Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, p. 1433–1443.

PENNINGTON J., SOCHER R. & MANNING C. D. (2014). Glove : Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, p. 1532–1543.

REN P., XIAO Y., CHANG X., HUANG P.-Y., LI Z., GUPTA B. B., CHEN X. & WANG X. (2021). A survey of deep active learning. *ACM Computing Surveys (CSUR)*, **54**(9), 1–40.

RINGGER E., MCCLANAHAN P., HAERTEL R., BUSBY G., CARMEN M., CARROLL J., SEPPI K. & LONSDALE D. (2007). Active learning for part-of-speech tagging : Accelerating corpus annotation. In *Proceedings of the Linguistic Annotation Workshop*, p. 101–108.

SANTOS B. N. D., MARCACINI R. M. & REZENDE S. O. (2021). Multi-domain aspect extraction using bidirectional encoder representations from transformers. *IEEE Access*, **9**, 91604–91613.

SCHEFFER T., DECOMAIN C. & WROBEL S. (2001). Active hidden markov models for information extraction. In *International Symposium on Intelligent Data Analysis*, p. 309–318 : Springer.

SCHRÖDER C. & NIEKLER A. (2020). A survey of active learning for text classification using deep neural networks. *arXiv preprint arXiv :2008.07267*.

SETTLES B. (2009). Active learning literature survey.

SETTLES B. & CRAVEN M. (2008). An analysis of active learning strategies for sequence labeling tasks. In *proceedings of the 2008 conference on empirical methods in natural language processing*, p. 1070–1079.

SHELMANOV A., LIVENTSEV V., KIREEV D., KHROMOV N., PANCHENKO A., FEDULOVA I. & DYLOV D. V. (2019). Active learning with deep pre-trained models for sequence tagging of clinical and biomedical texts. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, p. 482–489.

SHEN Y., YUN H., LIPTON Z. C., KRONROD Y. & ANANDKUMAR A. (2018). Deep active learning for named entity recognition. In *International Conference on Learning Representations*.

XU H., LIU B., SHU L. & YU P. S. (2018). Double embeddings and CNN-based sequence labeling for aspect extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 592–598.

Apprentissage profond pour l'estimation du quotient ouvert à partir du signal électroglottographique

Minh-Châu Nguyên Maximin Coavoux Solange Rossato
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG
minhchau.ntm@gmail.com, maximin.coavoux@univ-grenoble-alpes.fr,
Solange.Rossato@univ-grenoble-alpes.fr

MOTS-CLÉS : Électroglottographie, quotient ouvert glottique, apprentissage profond.

KEYWORDS: Electroglottography, glottal open quotient, deep learning.

1 Introduction

Contexte Les outils du Traitement Automatique des Langues (TAL) peuvent venir en appui à l'étude des langues rares pour les tâches de documentation fondamentale (telles que la transcription automatique de la parole, Adams *et al.*, 2018; Foley *et al.*, 2018). Dans cet article, nous cherchons à déterminer dans quelle mesure l'apprentissage machine peut également faciliter le traitement d'autres types de données expérimentales collectées sur ces langues (au sujet de l'éventail des techniques utilisées en phonétique, voir Vaissière *et al.* 2010). Spécifiquement, notre travail porte sur les enregistrements électroglottographiques (EGG). L'électroglottographie est une méthode non invasive pour estimer l'évolution de la surface d'accolement des plis vocaux au cours de la parole (Fabre, 1957; Fourcin *et al.*, 1995). Nous présentons ici les résultats d'expériences qui visent à automatiser l'estimation du quotient d'ouverture glottique (O_q) à partir du signal électroglottographique. Cette tâche est relativement incertaine en l'absence de vérification manuelle (Orlikoff, 1998; Herbst, 2020), et chronophage dans l'approche semi-automatique pratiquée dans plusieurs travaux (Michaud, 2004a; Recasens & Mira, 2013; Michaud *et al.*, 2015; Gao, 2015).

Objet de l'étude Le quotient d'ouverture glottique (O_q, unité : %) est le rapport entre la durée de la phase ouverte (entre une ouverture et la fermeture suivante) et celle du cycle glottique entier : O_q = (phase ouverte) / (phase ouverte+phase fermée). O_q est couramment considéré comme un paramètre lié au type de phonation : un O_q bas indique une phonation *pressée* ; un O_q moyen s'observe en voix modale ; et un O_q élevé est le signe d'une phonation fluide, à haut débit d'air (voix murmurée, voix soufflée). Ce paramètre peut être estimé au moyen du signal électroglottographique (Henrich *et al.*, 2004). La détection de pics positifs dans la dérivée première du signal EGG permet d'estimer l'instant de fermeture glottique, et un pic négatif unique et bien marqué entre deux pics positifs est considéré comme l'indice de l'instant d'ouverture de la glotte (figure 1).

L'estimation de l'O_q nécessite donc la détection de l'instant d'ouverture de la glotte. Une difficulté bien identifiée est qu'il est relativement courant que les pics d'ouverture ne soient pas uniques et bien marqués. Là où la mesure de la fréquence fondamentale (f_0) s'appuie sur la détection des pics de fermeture, qui dans la grande majorité des cas sont bien marqués, la détection des pics d'ouverture se heurte souvent à des difficultés dues à des pics imprécis. Parfois, aucun pic ne se détache clairement. Parfois, deux pics

Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain *(LIFT) et* Traitement Automatique des Langues *(TAL),*
Marseille, France, 14 au 15 novembre 2022 , pages 29–38,

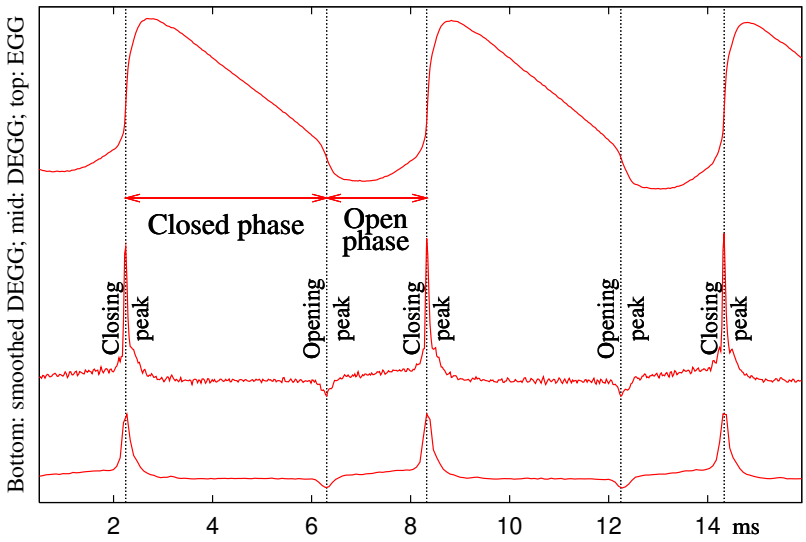


FIGURE 1 – Exemple de signaux EGG et dEGG avec indication de la fermeture et de l'ouverture de la glotte. Reproduit avec la permission de l'auteur, Alexis Michaud.

ou plus sont présents (pics multiples). La recherche des pics d'ouverture est particulièrement délicate dans le cas d'une phonation non modale, par exemple lorsque le voisement passe à la voix craquée (comme en figure 2). La vérification de l'O_q nécessite de poser des critères pour les situations particulières. Quand l'œil exercé peut-il, avec un bon degré de confiance, estimer la position de l'instant d'ouverture ? Quand faut-il renoncer à estimer le quotient ouvert ? La présente étude aborde cette question en entraînant un modèle neuronal sur un corpus annoté manuellement (par la première autrice du présent travail).

Corpus Le corpus utilisé provient d'un dialecte du muong (Vietnam). Cette langue possède un système tonal complexe au plan phonétique, combinant des contours de f_0 et des caractéristiques phonatoires. En particulier, dans le système de 5 + 2 tons [1], l'un des tons comporte un passage en voix craquée (Nguyen, 2021). Le signal EGG a été enregistré simultanément avec le signal acoustique pour 20 locuteurices (10 hommes, 10 femmes). Le corpus audio entier est en libre accès dans la collection Pangloss [2] (pour l'accès aux fichiers électroglottographiques, contacter la première autrice du présent travail). Le jeu de données utilisé ici est constitué de 12 ensembles de 5 mots monosyllabiques qui s'opposent par leur ton ("ensembles minimaux") parmi les syllabes sans occlusive finale et 3 paires minimales tonales parmi les syllabes à occlusive finale. Le corpus comprend des syllabes cibles à la fois isolées ("forme de citation") et dans une phrase-cadre de quatre mots (y compris le mot cible). La durée d'enregistrement de chaque locuteur est d'environ 15 minutes, donc environ 5 heures pour l'ensemble des données de 20 locuteurs. Il contient en tout près de 13 000 syllabes, ce qui représente 420 000 cycles glottiques.

Pour chaque cycle glottique, on dispose de la fréquence fondamentale (f_0) et de 5 valeurs de quotient ouvert. Les 4 premières sont le fruit d'une analyse automatique utilisant PeakDet (https://github.com/alexis-michaud/egg). Cet algorithme permet de choisir entre deux méthodes : détection des maxima sur la dérivée du signal EGG, ou méthode des barycentres, chacune s'appliquant avec ou sans lissage du signal (équivalent à un filtrage passe-bas). La 5^e valeur est le fruit d'une étape manuelle consistant à annoter chaque cycle glottique par une ou plusieurs des étiquettes suivantes :

1. Le système tonal du dialecte en question oppose 5 tons dans les syllabes sans occlusive finale (syllabes ouvertes sans consonne finale, et syllabes se terminant par une coda nasale) et 2 tons dans les syllabes à occlusive finale (/-p, -t, -k, -c/).
2. https://pangloss.cnrs.fr/ (Michaud *et al.*, 2016).

(0) pas d'O_q calculable, (1) et (2) choix de la méthode des maxima respectivement sans et avec lissage, (3) et (4) choix de la méthode des barycentres sans et avec lissage. La vérification revient ainsi, pour chaque cycle, à indiquer une double information : si l'estimation du quotient ouvert à partir du signal EGG paraît pratiquable pour le cycle en question ; et si tel est le cas, quelle méthode donne l'estimation la plus adéquate.

La figure 2 illustre un cas de signal dEGG bruité même après lissage. Il a été choisi d'exclure le cycle du milieu et le dernier (les 3^{e} et 5^{e} sur la fenêtre de cinq cycles mis en relief sur la figure). Le troisième comporte deux pics négatifs (ou deux "bosses" négatives), dont aucun n'est vraiment plus marqué que l'autre. Le cinquième présente un minimum clair, mais qui n'a pas une forme nette de pic. Pour les trois autres cycles, en présence d'un pic unique qui paraît bien visible, on retient ce pic sur le signal dEGG lissé (méthode 2).

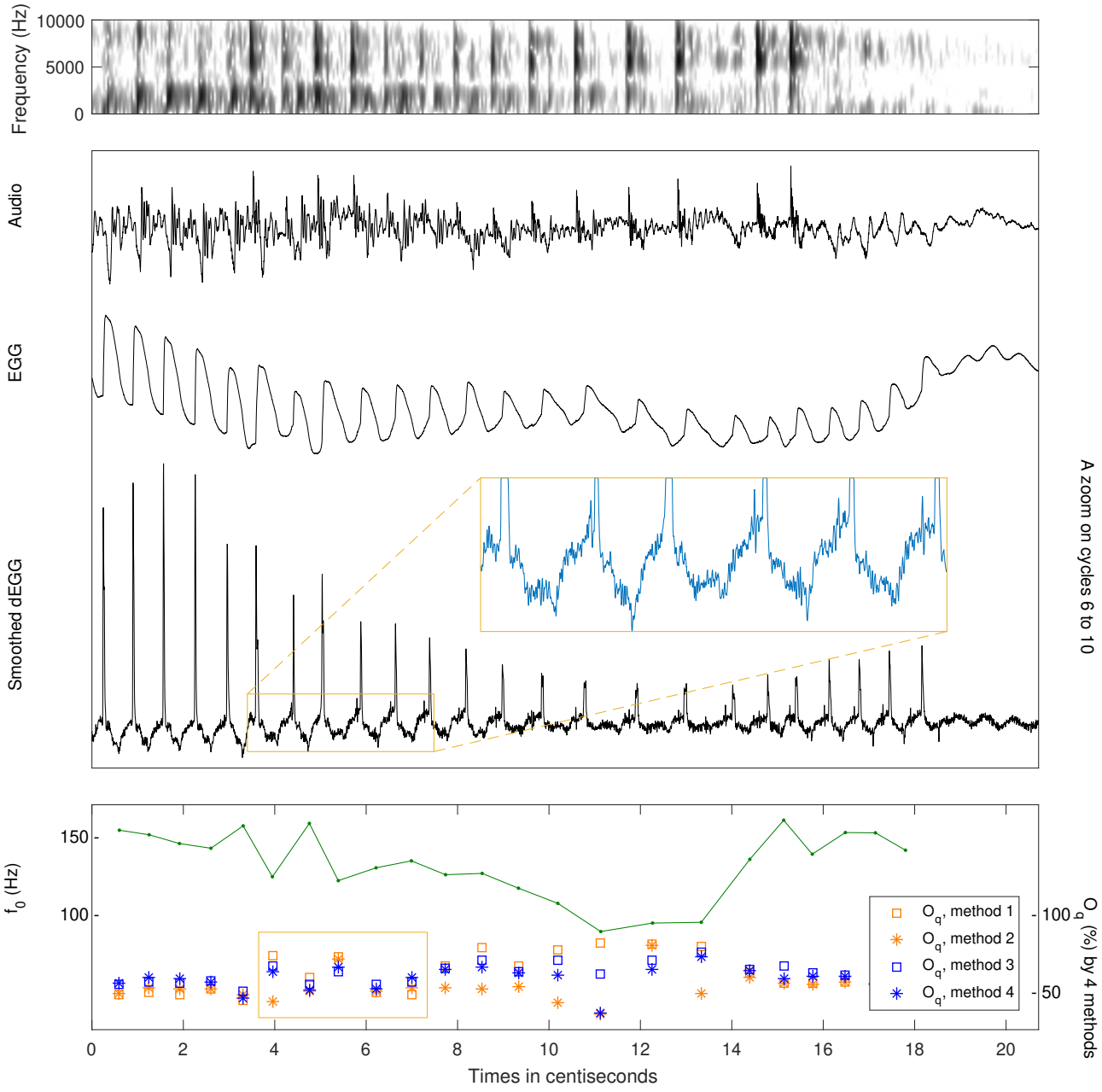


FIGURE 2 – Exemple de cas où la méthode 2 (minimum local sur le signal EGG lissé) a été retenue pour les cycles 1, 2 et 4 ; aucune valeur retenue pour les cycles 3 et 5. Données du locuteur F3, syllabe : /na4/ "tir à l'arc" (DOI : https://doi.org/10.24397/pangloss-0006761#W90).

Dans la figure 3, le signal est moins bruité, et la dérivée lissée est assez claire. Dans le 3^{e} des six cycles mis en valeur, un pic unique se détache. Pour les cycles 1, 2 et 4, le pic est globalement net mais bifurque

à son sommet, en forme de “fourche à deux dents”. Pour les cycles 5 et 6, il y a clairement deux pics, mais proches l’un de l’autre (la différence entre les estimations de O_q n’est que de l’ordre de 5% selon qu’on choisit le premier ou le second). Au vu de cette situation, on retient la méthode 4 : un barycentre entre les pics voisins (au sein du même cycle), pondéré par la hauteur des pics. Pour plus de détails sur l’algorithme de calcul du barycentre, nous renvoyons à la documentation du script PeakDet.

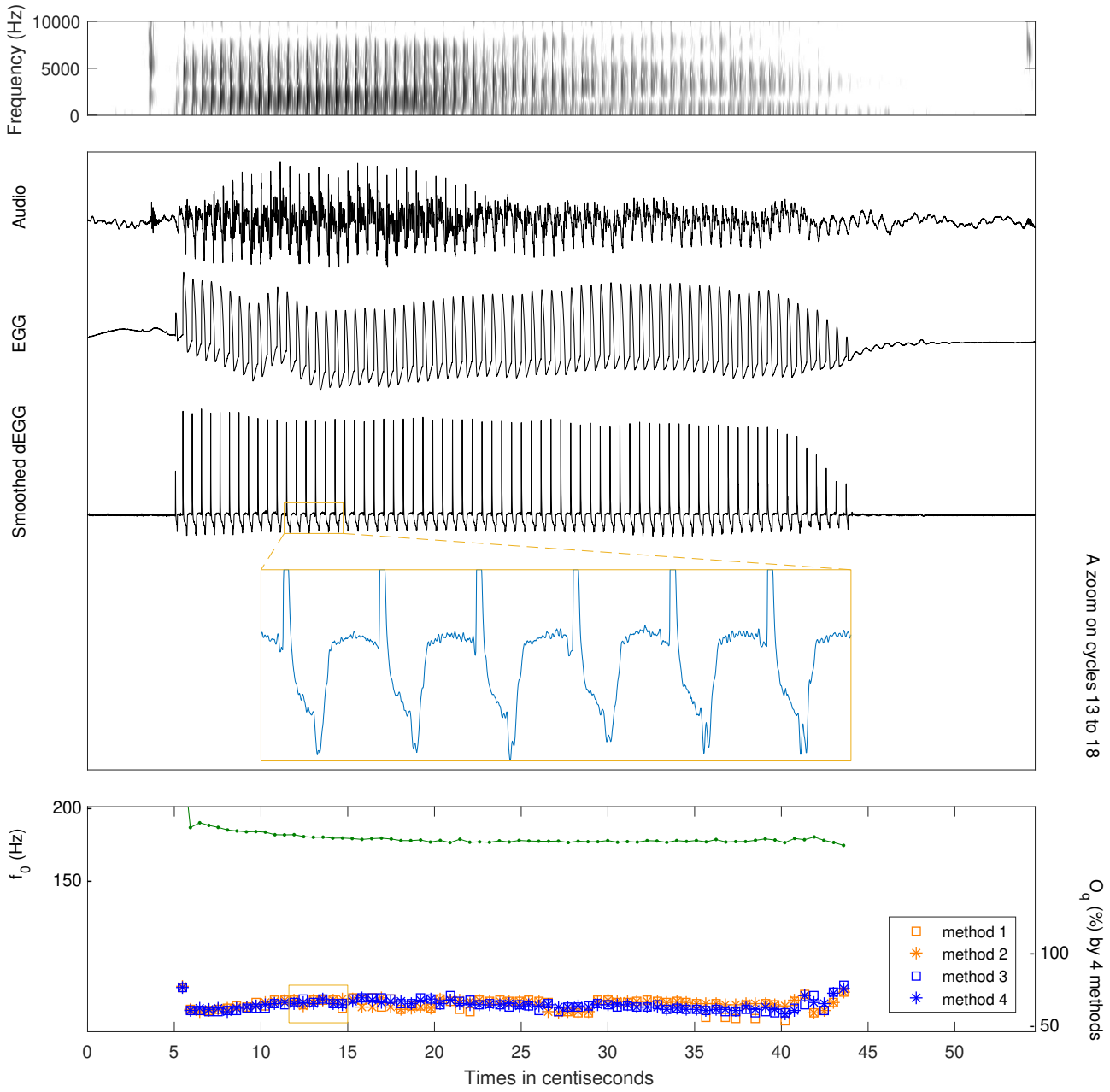


FIGURE 3 – Exemple de cas où l’option du barycentre des pics sur le signal dEGG lissé (méthode 4, représentée par des étoiles bleues) a été choisie. Données du locuteur F3, syllabe : /kaj1/ “cabbage” (DOI : https://doi.org/10.24397/pangloss-0006761#W107).

La méthode (3) n’est choisie que pour moins d’une centaine de cycles glottiques. En effet, le recours à un barycentre entre pics (méthodes 3 et 4) a lieu dans les cas où le pic n’est pas unique et bien marqué, or dans de tels cas, il serait paradoxal de ne pas recourir au lissage du signal (donc à la méthode 4 plutôt qu’à la méthode 3).

Le fruit des décisions ainsi réalisées sur l’ensemble du corpus est considéré comme une “vérité terrain” *(gold standard)* dans les expériences d’apprentissage neuronal exposées ci-dessous.

Section	Entraînement	Développement	Test	Corpus complet
Nombre de syllabes	9050	1913	2011	12974
Nombre de cycles glottiques	295753	62350	65727	423830
Distribution des étiquettes				
0 = aucun	69237 (23.41%)	14583 (23.39%)	14670 (22.32%)	98490 (23.24%)
1 = maxima sans lissage	60527 (20.47%)	13227 (21.21%)	13461 (20.48%)	87215 (20.58%)
2 = maxima avec lissage	137461 (46.48%)	29909 (47.97%)	30386 (46.23%)	197756 (46.66%)
3 = barycentre sans lissage	42603 (14.40%)	7870 (12.62%)	10102 (15.37%)	60575 (14.29%)
4 = barycentre avec lissage	93583 (31.64%)	18775 (30.11%)	21723 (33.05%)	134081 (31.64%)

TABLE 1 – Statistiques sur le corpus. Les méthodes implémentées par PeakDet donnant parfois les mêmes valeurs d'O_q, certains cycles ont plusieurs étiquettes correctes. Par conséquent, la somme des pourcentages des étiquettes peut dépasser 100%.

À la lumière des explications ainsi fournies au sujet du corpus de départ, nous pouvons maintenant passer à l'exposé de nos expérimentations. Elles ont pour but d'automatiser la décision manuelle concernant O_q, particulièrement chronophage. Pour le besoin des expérimentations, nous avons divisé le corpus en 3 sections : apprentissage (70%), validation (15%), test (15%). Nous présentons dans la table 1 quelques statistiques sur le corpus, dont la distribution des étiquettes. Nous observons ainsi que pour environ 23% du corpus, il n'y a pas d'O_q calculable. L'étiquette (2) est manuellement sélectionnée, seule ou avec d'autres méthodes, pour 46% des cycles, tandis que ce taux reste à 14% pour la méthode (3), montrant une répartition déséquilibrée des étiquettes.

2 Expérimentations

Modèle Nous avons implémenté un réseau de neurones basé sur un LSTM bidirectionnel destiné à prédire pour chaque cycle glottique, le résultat de l'annotation manuelle décrite dans la section précédente. Le réseau prend en entrée le signal EGG et éventuellement des informations additionnelles pour chaque cycle glottique : les chronocodes de chaque cycle (temps de début et de fin), la f_0, et les valeurs d'O_q estimées par les 4 méthodes implémentées par PeakDet. Pour représenter le signal EGG, nous utilisons des MFCC, avec une fenêtre de 6 ms glissante toutes les 2 ms (ces valeurs sont faibles par rapport aux valeurs habituellement utilisées en traitement de la parole, pour tenir compte de la granularité des représentations dont nous avons besoin). Nous obtenons une matrice $\mathbf{M^{(0)}}$ de taille $N \times F$ où N est le nombre de trames MFCC (c'est–à-dire la longueur du signal en millisecondes divisée par 2) et F est le nombre de traits MFCC extraits pour chaque trame. Ensuite, cette matrice est donnée en entrée à un réseau à propagation avant :

$$\mathbf{M}^{(1)} = \tanh(\mathbf{W}^{(1)} \cdot \mathrm{LayerNorm}(\mathbf{M}^{(0)}) + \mathbf{b}^{(1)}),$$

et contextualisée à l'aide d'un LSTM bidirectionnel :

$$\mathbf{M}^{(2)} = \text{bi-LSTM}(\mathbf{M}^{(1)}). \quad (1)$$

Pour représenter chaque cycle glottique c, nous utilisons la concaténation de 3 vecteurs $\mathbf{v}_c = [\mathbf{M}^{(2)}_{c_d}; \mathbf{M}^{(2)}_{c_f}; \mathbf{o}_c]$, où c_d et c_f sont les indicateurs temporels respectifs du début et de la fin du cycle, et $\mathbf{v}_c \in \mathbb{R}^5$ est un vecteur de traits additionnels contenant les 4 valeurs de quotient ouvert ainsi que la f_0 du

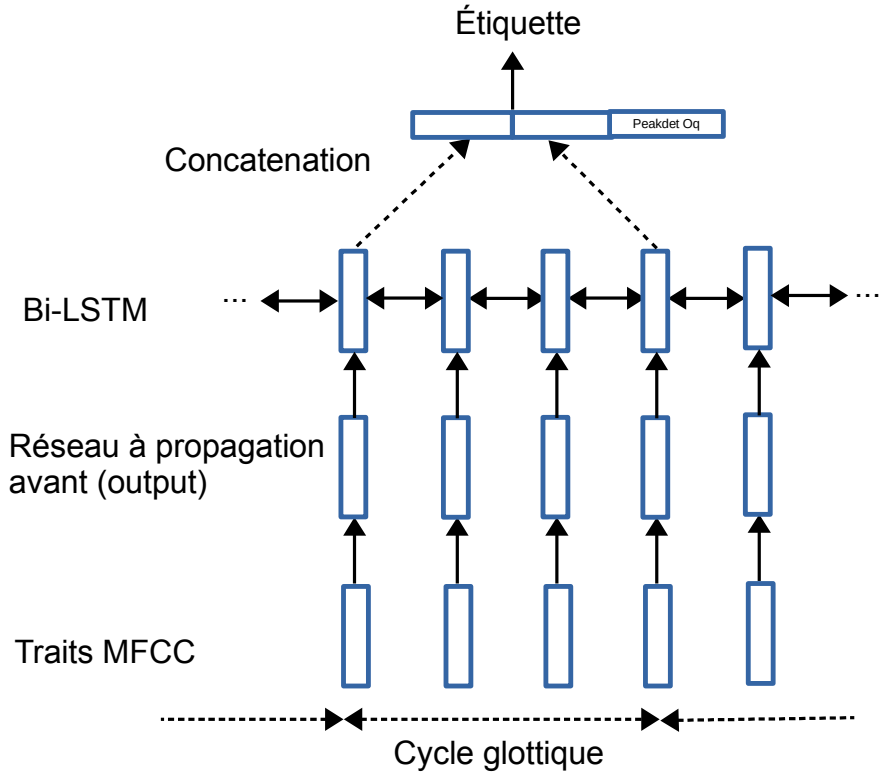


FIGURE 4 – Illustration du fonctionnement du réseau de neurones (centré sur un seul cycle glottique). En pratique, le bi-LSTM encode une syllabe complète.

cycle. Enfin, pour réaliser une prédiction pour un cycle, nous utilisons un second réseau à propagation avant :

$$\mathbf{P} = \text{Sigmoid}(\mathbf{W}^{(3)} \cdot \text{ReLU}(\mathbf{W}^{(2)} \cdot \text{LayerNorm}(\mathbf{v}_c) + \mathbf{b}^{(2)}) + \mathbf{b}^{(3)}),$$

où chaque $\mathbf{P} = [P(y_0 = 1|c), ..., P(y_4 = 1|c)]$ donne la probabilité est la probabilité que chaque étiquette soit correcte. Le réseau de neurones est illustré en figure 4.

Nous optimisons tous les paramètres du réseau à l'aide de l'algorithme Adam (Kingma & Ba, 2015), par maximisation des probabilités des étiquettes de référence. Lors de l'évaluation du modèle, nous considérons l'étiquette de plus forte probabilité comme prédiction du réseau. Les hyperparamètres du système sont les suivants :

- pour les MFCCs : taille des fenêtres (6ms), taille du pas entre chaque fenêtre (2ms), taille du contexte (2 trames de chaque côté) ;
- pour le réseau : dimension des couches cachées (128 unités pour les réseaux à propagation avant et 128 pour chaque direction du LSTM bidirectionnel) ;
- pour l'entraînement : algorithme d'optimisation (Adam), pas d'apprentissage (0.008 pour les modèles utilisant un signal, 0.001 pour le modèle sans signal), taille des *batches* (8), nombre d'époques.

Nous avons calibré les hyperparamètres lors d'expériences préliminaires (en particulier le pas d'apprentissage) sur le corpus de validation. Pour les expériences finales dont nous rapportons les résultats plus bas, nous entraînons chaque modèle pour 100 époques et sélectionnons l'époque qui maximise l'exactitude sur le corpus de validation avant évaluation finale sur le corpus de test.

Configurations expérimentales Nous cherchons à déterminer si l'utilisation du signal EGG permet d'améliorer la prédiction des étiquettes et s'il apporte une information supplémentaire aux valeurs d'O_q

Modèle	Validation			Test		
	Exactitude-3	Exactitude-2	Fscore-2	Exactitude-3	Exactitude-2	Fscore-2
Baseline (classe la plus fréquente)	48.9	76.6	0	46.2	77.7	0
(i) EGG + O_q + f_0	**63.6**	**86.1**	**91.0**	58.4	**85.8**	**91.0**
(ii) EGG	56.9	80.6	87.7	53.2	78.2	86.1
(iii) O_q + f_0	58.7	83.2	89.1	56.9	83.6	89.5
(iv) Audio + O_q + f_0	63.4	85.9	**91.0**	**59.4**	85.2	90.6
(v) Audio	57.1	78.9	86.6	51.5	79.2	87.3

TABLE 2 – Résultats finaux sur les sections de développement et de test (%).

pred ↓ gold→	0	1	1/2	2	2/3	2/3/4	2/4	3/4	4
0	61.77*	6.9	6.46	9.24	4.11	1.1	1.61	3.27	13.13
1	0	0*	0*	0.01	0	0	0	0	0
2	26.79	89.66	73.2*	72.67*	76.71*	74.88*	79.03*	60.23	50.35
3	0.04	0	0.29	0.22	0*	0.37*	0	0.46*	0.08
4	11.4	3.45	20.05	17.87	19.18	23.66*	19.35*	36.04*	36.44*

TABLE 3 – Matrice de confusion (corpus de test) pour le modèle (iv). Les valeurs données sont des pourcentages calculées sur les étiquettes gold. Les prédictions correctes sont indiquées par *.

et f_0 estimés par PeakDet. Par ailleurs, nous cherchons à savoir si le signal EGG est plus informatif que le simple signal audio. Nous évaluons ainsi plusieurs types d'entrée différents pour étudier le comportement du modèle :

— (i) signal EGG + O_q estimé par PeakDet + f_0 : $[\mathbf{M}_{c_d}^{(2)}; \mathbf{M}_{c_f}^{(2)}; \mathbf{o}_c]$;
— (ii) signal EGG seul : $[\mathbf{M}_{c_d}^{(2)}; \mathbf{M}_{c_f}^{(2)}]$;
— (iii) O_q et f_0 estimés par PeakDet : $[\mathbf{o}_c]$;
— (iv) comme (i) mais en utilisant le signal audio à la place du signal EGG ;
— (v) comme (ii) mais en utilisant le signal audio à la place du signal EGG.

Résultats Nous rapportons les résultats dans le tableau 2. Les métriques d'évaluation sont : l'exactitude à 3 classes (0 vs 1-2 vs 3-4), l'exactitude et le Fscore à 2 classes (étiquette 0 vs autre étiquette), pour les 5 modèles, ainsi qu'une baseline choisissant systématiquement la classe la plus fréquente. Les valeurs s'élèvent (modestement) au-dessus de la baseline, signe qu'il y a eu apprentissage. Les résultats obtenus avec le signal EGG seul en entrée (ii) montrent que celui-ci n'est pas exploité à plein, puisque les résultats sont moins bons qu'en (iii) lorsque l'on utilise uniquement les 4 méthodes de PeakDet et la f_0. Lorsque les deux informations sont présentes (i), les performances sont meilleures. Par ailleurs, les valeurs pour les entrées (i) et (iv) sont quasi-identiques, indiquant que l'information du signal EGG et celle du signal acoustique sont sensiblement équivalentes pour le système, en complément des estimations de O_q par PeakDet.

Matrice de confusion Nous présentons la matrice de confusion du modèle (iv), qui a obtenu les meilleurs résultats dans le tableau 3. Chaque colonne représente une combinaison d'étiquettes vue dans les données de référence (par exemple, la colonne 3/4 représente les cycles glottiques où les méthodes PeakDet 3 et 4 ont donné la bonne valeur. On observe que les étiquettes 1 et 3, qui sont les moins présentes dans les données (et donnent souvent le même résultat respectivement que les méthodes 2 et 4), ne sont quasiment

jamais prédites, ce qui montre que le système ne discrimine pas les classes 1 et 2 d'une part et les classes 2 et 3 d'autre part. Les erreurs les plus fréquentes sont la difficulté à prédire les classes 3, 4 et 0, où le modèle se replie sur la classe la plus fréquente (2).

3 Perspectives

Études des types de phonation Le travail pluridisciplinaire décrit ici a des implications concernant l'étude phonétique des types de phonation. En effet, l'ensemble de résultats obtenu ici met en lumière le fait que le processus qui consiste à évaluer la fiabilité de l'estimation du quotient ouvert au vu du signal constitue une tâche non triviale. Cela fournit l'occasion de revenir sur la question fondamentale de ce que reflète le quotient ouvert glottique, et de son interprétation. Le quotient ouvert constitue une projection linéaire de phénomènes non linéaires qui entrent en jeu dans la phonation, et ne saurait donc, à l'évidence, constituer par lui-même un descripteur suffisant des divers types de phonation (voix murmurée, voix pressée, voix craquée ; parmi les référentiels couramment employés, voir notamment Laver, 1980). Spécifiquement dans le cas de la voix craquée, illustrée ci-dessus par la figure 2, on observe une bonne corrélation entre la présence de voix craquée et l'information de pente spectrale reflétée dans le spectrogramme (une intensité plus forte dans la moitié supérieure du spectrogramme, de 5 à 10 kHz, que dans la moitié inférieure : cycles 2, 6, 8, 9, 13, 14, 15, 17, 18). Tandis que le quotient ouvert glottique pour les cycles en question ne montre pas de valeurs exceptionnellement basses. Le signal audio est donc ici un meilleur outil que le quotient ouvert pour détecter la voix craquée. Le signal EGG contient d'autres informations, à commencer par la fréquence fondamentale, qui fournissent des indications plus claires que O_q concernant le type phonatoire.

Au plan acoustique, il est connu que le quotient ouvert n'est ni le seul, ni le plus important parmi les paramètres de source glottique. Le fait qu'on puisse en obtenir une estimation à partir du signal EGG a sans doute amené à lui accorder une importance particulière, en comparaison par exemple du quotient de vitesse (*speed quotient*), qui, lui, n'est pas aisément accessible à une estimation. Ainsi, la hauteur du pic positif sur la dérivée (correspondant à l'instant de fermeture glottique en début et fin de cycle), DECPA (pour Derivative-Electroglottographic Closure Peak Amplitude, Michaud, 2004b), ne permet hélas pas d'estimer directement le quotient de vitesse. Clairement, O_q gagnerait donc à être intégré à des expériences d'apprentissage machine dans lesquelles il serait intégré au sein d'un ensemble élargi de paramètres acoustiques, de façon à caractériser divers types de phonation d'une façon à la fois objective et complète.

Perspectives pour la suite du travail Dans la suite de ce travail, nous prévoyons de développer et d'évaluer des modèles de régression permettant de prédire directement la variable O_q, au lieu de la prédire indirectement via une tâche de classification. Par ailleurs, nous prévoyons d'évaluer le remplacement des traits MFCC par des traits extraits par des modèles acoustiques préentraînés (Conneau *et al.*, 2020).

Remerciements

Les travaux présentés dans cet article sont financés 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). Nous remercions les relecteurices anonymes pour leurs commentaires. Merci à Alexis Michaud pour de nombreuses discussions et sa relecture attentive.

Références

ADAMS O., COHN T., NEUBIG G., CRUZ H., BIRD S. & MICHAUD A. (2018). Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *LREC 2018 (Language Resources and Evaluation Conference)*, p. 3356–3365. HAL : halshs-01709648.

CONNEAU A., BAEVSKI A., COLLOBERT R., MOHAMED A. & AULI M. (2020). Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv :2006.13979*.

FABRE P. (1957). Un procédé électrique percutané d'inscription de l'accolement glottique au cours de la phonation : glottographie de haute fréquence. *Bulletin de l'Académie Nationale de Médecine*, **141**, 66–69.

FOLEY B., ARNOLD J. T., COTO-SOLANO R., DURANTIN G., ELLISON T. M., VAN ESCH D., HEATH S., KRATOCHVIL F., MAXWELL-SMITH Z., NASH D. *et al.* (2018). Building speech recognition systems for language documentation : The CoEDL Endangered Language Pipeline and Inference System (ELPIS). In *SLTU*, p. 205–209.

FOURCIN A., ABBERTON E., MILLER D. & HOWELLS D. (1995). Laryngograph : speech pattern element tools for therapy, training and assessment. *European Journal of Disorders of Communication*, **30**(2), 101–115.

GAO J. (2015). *Interdependence between tones, segments and phonation types in Shanghai Chinese*. Ph.D., Université Sorbonne Nouvelle.

HENRICH N., D'ALESSANDRO C., DOVAL B. & CASTELLENGO M. (2004). On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation. *The Journal of the Acoustical Society of America*, **115**(3), 1321–1332.

HERBST C. T. (2020). Electroglottography–an update. *Journal of Voice*, **34**(4), 503–526.

KINGMA D. P. & BA J. (2015). Adam : A method for stochastic optimization. *CoRR*, **abs/1412.6980**.

LAVER J. (1980). *The phonetic description of voice quality*. Cambridge : Cambridge University Press.

MICHAUD A. (2004a). Final consonants and glottalization : new perspectives from Hanoi Vietnamese. *Phonetica*, **61**(2-3), 119–146.

MICHAUD A. (2004b). A Measurement from Electroglottography : DECPA, and its Application in Prosody. In B. BEL & I. MARLIEN, Éds., *Speech Prosody 2004*, p. 633–636, Nara, Japan.

MICHAUD A., GUILLAUME S., JACQUES G., MAC Đ.-K., JACOBSON M., PHAM T.-H. & 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 de *Actes de la conférence conjointe JEP-TALN-RECITAL 2016, volume 1 : Journées d'Etude de la Parole*, p. 155–163, Paris, France : Association Francophone de la Communication Parlée. HAL : halshs-01341631.

MICHAUD A., VAISSIÈRE J. & NGUYÊN M.-C. (2015). Phonetic insights into a simple level-tone system : 'careful' vs. 'impatient' realizations of Naxi High, Mid and Low tones. In *ICPhS XVIII (18th International Congress of Phonetic Sciences)*. HAL : halshs-01148765.

NGUYEN M.-C. (2021). *Glottalization, tonal contrasts and intonation : an experimental study of the Kim Thuong dialect of Muong*. thèse de doctorat, Université de la Sorbonne nouvelle - Paris III. HAL : tel-03652510.

ORLIKOFF R. F. (1998). Scrambled EGG : The uses and abuses of electroglottography. *Phonoscope*, **1**(1), 37–53.

RECASENS D. & MIRA M. (2013). Voicing assimilation in Catalan three-consonant clusters. *Journal of Phonetics*, **41**(3-4), 264–280.

VAISSIÈRE J., HONDA K., AMELOT A., MAEDA S. & CREVIER-BUCHMAN L. (2010). Multisensor platform for speech physiology research in a phonetics laboratory. *Journal of the Phonetic Society of Japan*, **14**(2), 65–77.

Automatic Legal Document Analysis (A.L.D.A.) : Plateforme d'analyse automatique de documents juridiques

Ying ZHANG[1] Matthieu PETIT GUILLAUME[1] Aurélien KRAUTH[1]
(1) Leviatan, 725 Boulevard Robert Barrier, 73100 Aix-les-Bains, France
y.zhang@leviatan.fr, matthieu@leviatan.fr, aurelien@leviatan.fr

RÉSUMÉ

Dans le cadre d'un besoin industriel, notre partenaire gère une grande quantité de contrats. Il consacre chaque année de nombreuses ressources humaines et matérielles afin de gérer ces documents. Il s'agit ici d'analyser ces documents, d'en extraire un ensemble d'informations, et de saisir/stocker des informations qualifiées dans une feuille de calcul, appelée feuille d'analyse. Dans cet article, nous proposons une nouvelle plateforme A.L.D.A. (Automatic Legal Document Analysis) permettant l'analyse automatique de documents juridiques. Cette plateforme a été implémentée en 3 modules indépendants : (1) ALDA-Parser pour uniformiser les documents juridiques, (2) ALDA-Classifier pour classifier les documents, et (3) ALDA-MRC pour générer des réponses automatiques.

ABSTRACT

Automatic Legal Document Analysis (A.L.D.A.): Platform for automatic analysis of legal documents

As part of an industrial need, our partner manages a large number of contracts. Each year, he devotes numerous human and material resources to managing these documents. It involves analyzing these documents, extracting a set of information from them, and entering/storing qualified information in a spreadsheet, called an analysis sheet. In this article, we propose a new A.L.D.A. (Automatic Legal Document Analysis) allowing the automatic analysis of legal documents. This platform has been implemented in 3 independent modules: (1) ALDA-Parser to standardize legal documents, (2) ALDA-Classifier to classify documents, and (3) ALDA-MRC to generate automatic responses.

MOTS-CLÉS : Analyse automatique des documents juridiques, Compréhension automatique de texte, Système de questions-réponses, Classification des textes

KEYWORDS: Automatic legal document analysis, Machine reading comprehension, Question answering system, Classification of texts

Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain *(LIFT) et* Traitement Automatique des Langues *(TAL),*
Marseille, France, 14 au 15 novembre 2022 , pages 39–48,

1 Introduction du système A.L.D.A.

Dans le cadre d'un besoin industriel, nous avons stocké une grande quantité de documents juridiques. Ces documents sont divisés en plusieurs types, tels que des « baux commerciaux », « contrats de transports », et « contrats d'énergies » etc...

Notre partenaire consacre chaque année de nombreuses ressources humaines et matérielles afin de gérer ces documents. Il s'agit d'analyser ces documents, d'en extraire un ensemble d'informations, et de saisir/stocker des informations qualifiées dans une feuille de calcul, appelée feuille d'analyse.

Dans cet article, nous proposons une nouvelle plateforme A.L.D.A. (Automatic Legal Document Analysis) permettant l'analyse automatique de documents juridiques.

Pour chaque type de document, nous possédons une liste d'informations à analyser. Cette liste est prédéfinie par l'utilisateur. Par exemple, pour le type de document « bail commercial » nous pouvons avoir les questions suivantes :

- Quelle surface a été louée ?
- Quelle est la date de fin du bail ?
- Quel est le délai de préavis ?

Pour le type « contrat de transport », nous pouvons avoir les questions :

- Quel est le lieu de départ ?
- Quel est le lieu d'arrivée ?

L'utilisateur peut ainsi ajouter, supprimer ou mettre à jour ces questions à tout moment via une interface homme machine (IHM).

L'objectif du système est, après téléversement de documents juridiques par l'utilisateur via l'IHM, la mise en place d'un système de numérisation automatique de documents (extraction et structuration des textes des documents), la classification automatique du type du document, la génération automatique des réponses en fonction des questions prédéfinies par l'utilisateur, la restitution par affichage dans une IHM, et la fourniture automatique d'une feuille d'analyse structurées à télécharger.

Dans cet article, nous ne présentons pas la réalisation de l'interface homme machine ni la gestion des utilisateurs. Par conséquent, la plateforme A.L.D.A. est divisée en trois modules indépendants : ALDA-Parser (structuration des textes des documents), ALDA-Classifier (classification des documents), ALDA-MRC (analyse des informations pour un ensemble de questions prédéfinies).

2 ALDA-Parser

Les documents juridiques sont stockés dans des formats hétérogènes. Il existe des documents docx, des images (JPEG, PNG etc…), des PDF cryptés, des PDF éditables et des PDF scannés. La rotation des images et des scans n'est pas souvent correcte. Par exemple, la première page pourrait être tournée de 90 degrés dans le sens antihoraire puis la deuxième page pourrait être tournée de 45 degrés dans le sens horaire.

L'objectif du module ALDA-Parser est le traitement, l'uniformisation et l'unification de tous les documents afin de générer un texte brut pour chaque contrat, l'identification de la rotation de chaque page et la conversion du document en HTML éditable et en PDF éditable. Les documents originaux téléversés sont stockés dans le cloud Microsoft Azure DataLake. Les informations extraites sont stockées dans le Microsoft Azure Cosmos DB.

Nous avons intégré plusieurs logiciels en source ouverte dans ce module, tels que Lowriter (Damien, 2019), qpdf (Berkenbilt, 2005), pdf2image (Belval, 2019), ocrmypdf (Barlow, 2022), pdf2htmlEX (Wang, 2016), tesseract (Smith, 2007).

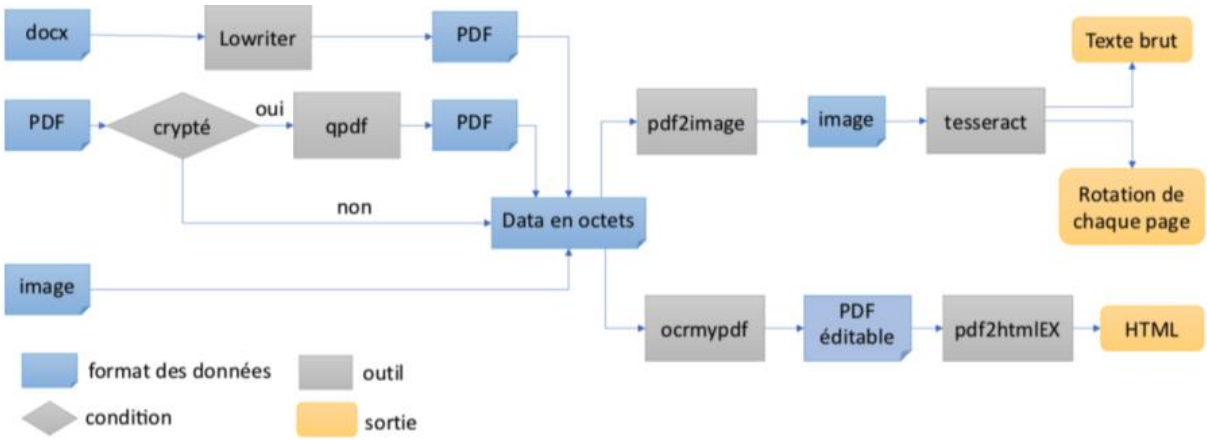


FIGURE 1 : Flux de travail d'ALDA-Parser

3 ALDA-Classifier

L'objectif du module ALDA-Classifier est la classification de tous les documents juridiques, afin d'en retrouver le type et les informations à analyser (les questions prédéfinies par l'utilisateur du système A.L.D.A.).

3.1 État de l'art

La classification de texte est un sujet très important dans le domaine du TALN (Traitement Automatique du Langage Naturel). Des recherches récentes sur l'utilisation des modèles de langue contextualisés pré-entraînés avec des architectures à base de transformer (Vaswani et al., 2017) ont obtenu un grand succès dans de nombreuses tâches de TALN.

Nous essayons de réaliser un système de classification des documents intégrant une nouvelle méthode et des technologies de pointe : avec un petit jeu de données annoté, nous ajustons (fine-tuning) le modèle pré-entraîné afin d'améliorer la précision des résultats de classification et de fournir une solution plus rapide et plus économique.

CamemBERT (Martin et al., 2020) est le modèle pré-entraîné le plus approprié pour nos expérimentations en prenant en compte la disponibilité du code source et la qualité du modèle pré-entraîné. Par conséquent, nous utilisons CamemBERT comme point de départ pour développer notre solution.

3.2 Problématiques

Afin d'entraîner ou d'ajuster un modèle, une bonne qualité et une quantité correcte de données est nécessaire. En début de projet, nous recevons des documents juridiques hétérogènes sans aucune donnée annotée.

Grâce au module ALDA-Parser et le développement d'une interface d'annotation, nous avons pu récupérer les données annotées au fur et à mesure, avec une quantité pas assez importante.

La limite ne concerne pas seulement la quantité de données, mais également le nombre de types de documents inconnus et les modèles variés. En effet, le système A.L.D.A. est installé chez les différents clients finaux. Par exemple, avec un client d'enseigne de la grande distribution, nous observons des types de contrats « baux commerciaux » et « contrats de location de voiture ». Avec une entreprise d'énergie, nous observons d'autres types, tels que « contrats d'énergie gaz » et « contrats d'énergie électricité ». Nous ne pouvons pas entraîner un modèle adapté à toutes les utilisations, nous ne connaissons par ailleurs pas tous les types de documents. Il est donc impossible de déterminer le nombre de classes.

Un autre besoin concret représente la possibilité d'ajout d'un ou plusieurs types de contrat de façon dynamique. En d'autres termes, lors du déploiement d'un système, il peut n'y avoir que zéro ou un type de contrats, mais au fil du temps, il peut y en avoir plusieurs.

Afin de résoudre l'ensemble de ces problèmes, nous avons conçu un système intégrant un (ré)entraînement incrémental basé sur un modèle pré-entraîné.

3.3 Travaux réalisés

3.3.1 Structure de modèle

Dans notre solution, nous proposons un système de classification de documents intégrant une nouvelle méthode et des technologies de pointe en deux parties.

La première partie utilise le modèle CamemBERT (Martin et al., 2020) fine-tuné comme modèle d'extraction de caractéristiques. Nous respectons les consignes indiquées dans (Sun et al., 2019) afin d'ajuster cette première partie d'ALDA-Classifier. Les sorties de ce modèle ajusté sont des vecteurs de caractéristiques des textes bruts. La sortie est également l'entrée de la deuxième partie de notre modèle. Afin d'être plus précis, toutes les sorties du dernier encodeur de transformer CamemBERT ajusté, sont utilisées de manière à être alimentées en entrée d'un autre réseau neuronal récurrent (RNN) bidirectionnel (Bi-LSTM) (Liang & Zhang, 2016), permettant la résolution de nombreux problèmes de dépendance à long terme du modèle RNN classique. Voir FIGURE 2.

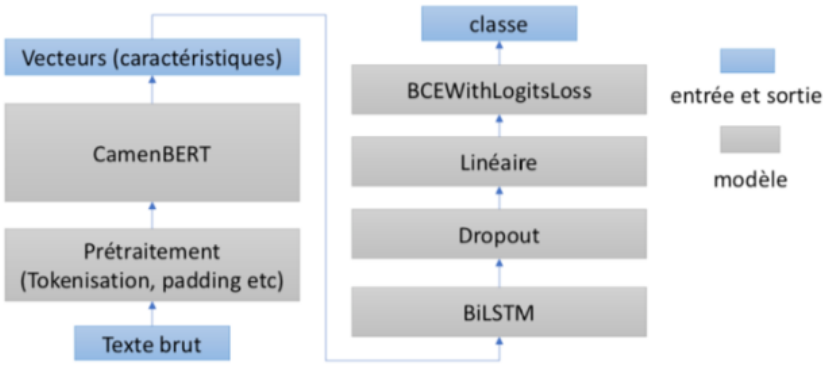


FIGURE 2 : Structure d'ALDA-Classifier

La longueur d'un contrat peut atteindre une centaine de pages, nous avons récupéré les 512 premiers tokens comme entrée du modèle. Nos expériences ont prouvé que cette longueur est suffisante pour que ALDA-Classifier reconnaisse le type du document (voir la section 3.3.3).

Nous pouvons souligner que la fonction loss de ALDA-Classifier est « BCEWithLogitsLoss »[1]. Dans la première version du ALDA-Classifier, nous avons utilisé « softmax » comme fonction d'activation de la dernière couche. C'est-à-dire que la réalisation en PyTorch de la fonction loss est « CrossEntropyLoss »[2]. En raison de l'appartenance d'un contrat à un seul et même type.

Notre partenaire a appliqué une nouvelle exigence après premières expériences d'utilisation : si le modèle a été entraîné sur des baux commerciaux et contrats d'énergie, pour un document d'un

[1] https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html
[2] https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html

nouveau type, ALDA-Classifier devrait être capable de prédire ce document qui n'est ni un bail commercial, ni un contrat d'énergie.

En raison de cette demande, il ne s'agit plus d'un problème « multi-classe », mais plutôt d'un problème de « multi-label » (Goyal, 2021). Nous avons remplacé la fonction « CrossEntropyLoss » par la fonction « BCEWithLogitsLoss ».

Avec la fonction « BCEWithLogitsLoss », ALDA-Classifier propose tous les types dans le cas ou la possibilité de ce type est supérieure ou égale à 0.5. En revanche, un contrat ne peut pas être attaché à plusieurs types. Nous stockons uniquement le type avec le score le plus élevé dans la base de données. Si nous obtenons plusieurs scores élevés et similaires, le système envoie une alarme sur l'interface afin de demander une vérification humaine.

3.3.2 Entraînement incrémental

ALDA-Classifier est une API qui contient deux fonctionnalités : « training » et « prediciton ». Cette API est installée chez les clients finaux. Chaque système déployé se connecte à une base de données (Cosmos DB) (Sahay, 2020) dédiée. Nous avons créé un ensemble de règles. Lorsque les conditions sont remplies, une requête « training » sera activée. Le résultat de nouveaux entraînements sera automatiquement chargé par le système.

Ce module a été implémenté en PyTorch (Paszke et al., 2017).

3.3.3 Résultat préliminaire

En raison de certaines restrictions commerciales, nous avons testé un millier de contrats en 3 types (« bail commercial », « contrat de transport » et « contrat d'énergie »). Nous avons utilisé 80% de contrats pour l'entraînement, et 20% de contrats pour le test. Avec un seul epoch, la précision du jeu de données test est 0.942, le macro F1 score est 0.816.

4 ALDA-MRC

4.1 État de l'art

Le système de questions/réponses dispose de divers mécanismes de mise en œuvre, tels que la recherche d'informations (IR : Information Retrieval), KB : Knowledge Base, KG : Knowledge Graph, etc. Un système de réponse aux questions relativement complet est souvent une combinaison de plusieurs mécanismes. Nous avons particulièrement réalisé une étude sur la compréhension de la lecture automatique (MRC : Machine Reading Comprehension).

Les modèles les plus performants exposent deux stratégies : 1. le modèle de langue contextualisé pré-entraîné tel que BERT (Devlin et al., 2019), ELMo (Peters et al., 2018), ou 2. le modèle d'encodage avancé tel que QANet (Yu et al., 2018).

Après nos études sur ces deux stratégies, nous utilisons CamemBERT comme le modèle de langue contextualisé pré-entraîné afin de commencer notre expérimentation.

4.2 Jeux de données

En français, 3 jeux de données sont disponibles et en source ouverte. Il s'agit du projet Piaf (Bras, 2019), du projet FQuAD (D'Hoffschmidt et al., 2020) et du projet French-SQuAD (Kabbadj, 2018). En plus de cela, notre partenaire a également une petite quantité de données privées extraites grâce aux feuilles d'analyse et aux documents connexes.

4.3 Travaux réalisés

4.3.1 Fine-tuning d'un MRC basé sur CamemBERT

Dans (Devlin et al., 2019), l'équipe Google AI Language a bien indiqué la procédure afin d'ajuster un système MRC basé sur BERT en utilisant les jeux de données SQuAD v1.1 et v2.0. Dans ce document nous n'entrerons pas dans les détails. Nous avons utilisé la même méthode afin d'entraîner notre MRC français basé sur CamemBERT et en utilisant les jeux de données français.

Pour le fine-tuning d'un MRC basé sur CamemBERT, nous ne modifions pas l'architecture d'origine de CamemBERT, nous avons seulement modifié les couches d'entrée et de sortie. *[CLS]* est un symbole spécial ajouté devant chaque exemple d'entrée, et *[SEP]* est un token de séparation spécial (par exemple, séparant les questions/réponses). Nous avons principalement conservé les valeurs par défaut des hyper-paramètres proposées par Transformer (Huggingface, 2020).

Dans notre système A.L.D.A., il y a une notion de « projet ». Chaque projet correspond à un ensemble de documents : généralement un contrat et plusieurs avenants. Pour certaines questions posées, par exemple, « Quelle surface a été louée ? » dans un projet de type « bail commercial », nous devons trouver une meilleure solution dans un ensemble de documents de ce projet. Si plusieurs solutions présentent des scores élevés, nous proposons la solution la plus récente selon la date de signature du document. Par exemple, si nous avons trouvé une première réponse dans le contrat signé en 2010, une deuxième réponse dans l'avenant signé en 2020. Nous proposons la deuxième réponse afin de générer la feuille d'analyse. Pour l'identification d'un projet et de la date de signature, l'utilisateur du système A.L.D.A. doit respecter une règle de nomenclature pour tous les documents téléversés.

Ce module a été implémenté en PyTorch.

4.3.2 Résultat préliminaire

Le F1-score atteint en moyenne 80.6% sur un ensemble de jeux de données. Ce modèle est un modèle pré-entraîné et générique. Contrairement au module ALDA-Classifier, nous n'avons pas recours à des entraînements incrémentaux pour ALDA-MRC.

5 Conclusions et perspectives du A.L.D.A.

A.L.D.A. a été mis en service en 2021 et a permis l'analyse de plus de 500 projets. Le partenaire industriel est satisfait de la qualité des feuilles d'analyses générées automatiquement par le système.

Actuellement, le système A.L.D.A. ne permet pas de traiter les données tabulées. À long terme, nous voulons intégrer un quatrième module ALDA-FormAnalyzer. Pour la conception de ce module, nous avons été inspirés par le service AWS Textract (AWS, 2020).

Nous avons prévu de séparer ce module en 2 parties : 1. un réseau de neurones permettant la détection de formulaires, 2. un parser permettant d'aligner et de traduire les données tabulées en langage naturel. Pour la première partie, nous avons commencé à annoter les données en utilisant le logiciel Labelimg (Nelson, 2020), ensuite nous avons prévu de tester Yolo3 (Redmon et al., 2018) et Tensorflow Object Detection (Krishna Sai & Sasikala, 2019) pour démarrer la recherche et le développement de ce module.

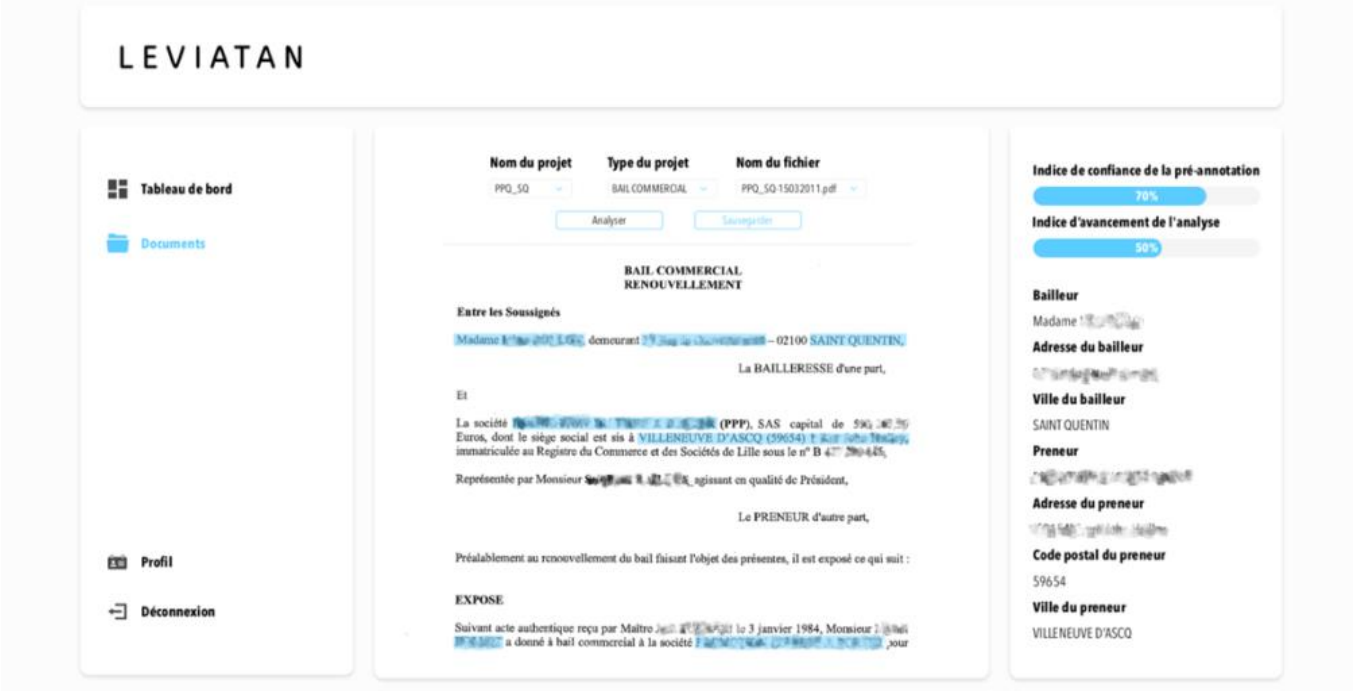


FIGURE 3 : Interface d'affichage d'un contrat (informations personnelles floutées par effet mosaïque)

Références

AWS. (2020). *Amazon Textract Developer Guide*. Available at: https://docs.aws.amazon.com/pdfs/textract/latest/dg/textract-dg.pdf

BARLOW, J. R. (2022). *OCRmyPDF documentation*. Available at: https://ocrmypdf.readthedocs.io/en/latest/

BELVAL, E. (2019). *pdf2image's documentation*. Available at: https://pdf2image.readthedocs.io/en/latest/index.html

BERKENBILT, J. (2005). *QPDF documentation*. Available at: https://qpdf.readthedocs.io/en/stable/

BRAS, M. (2019). *Piaf.* Available at: https://www.etalab.gouv.fr/ia-decouvrez-et-participez-au-projet-piaf-pour-des-ia-francophones

D'HOFFSCHMIDT, M., VIDAL, M., BELBLIDIA, W., & BRENDLÉ, T. (2020). FQuAD: French question answering dataset. DOI : https://doi.org/10.18653/v1/2020.findings-emnlp.107

DAMIEN, A. (2019). *Lowriter, convertir des documents en PDF à partir du terminal*. Available at: https://ubunlog.com/fr/lowriter-documentos-a-pdf-terminal/

DEVLIN, J., CHANG, M. W., LEE, K., & TOUTANOVA, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*.

GOYAL, C. (2021). Demystifying the difference between multi-class and multi-label classification problem statements in deep learning. *Data Science Blogathon - 10*.

HUGGINGFACE. (2020). *Fine-tuning BERT on SQuAD1.0*. Available at: https://huggingface.co/transformers/v2.8.0/examples.html#squad

KABBADJ, A. (2018). *Something new in French Text Mining and Information Extraction (Universal Chatbot): Largest Q&A French training dataset (110 000+)*.

KRISHNA SAI, B. N., & SASIKALA, T. (2019). Object Detection and Count of Objects in Image using Tensor Flow Object Detection API. *Proceedings of the 2nd International Conference on Smart Systems and Inventive Technology, ICSSIT 2019*. DOI : https://doi.org/10.1109/ICSSIT46314.2019.8987942

LIANG, D., & ZHANG, Y. (2016). *AC-BLSTM: Asymmetric Convolutional Bidirectional LSTM Networks for Text Classification*. DOI : https://doi.org/10.48550/arXiv.1611.01884

MARTIN, L., MULLER, B., SUAREZ, P. J. O., DUPONT, Y., ROMARY, L., DE LA CLERGERIE, É. V., SEDDAH, D., & SAGOT, B. (2020). CamemBERT: A tasty French language model. DOI : https://doi.org/10.18653/v1/2020.acl-main.645

NELSON, J. (2020). *LabelImg for Labeling Object Detection Data*. Roboflow. Available at: https://blog.roboflow.com/labelimg/

PASZKE, A., GROSS, S., CHINTALA, S., CHANAN, G., YANG, E., & ... (2017). Automatic differentiation in pytorch. In *NIPS-W*. https://openreview.net/forum?id=BJJsrmfCZ

PETERS, M. E., NEUMANN, M., IYYER, M., GARDNER, M., CLARK, C., LEE, K., & ZETTLEMOYER, L. (2018). Deep contextualized word representations. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*. DOI : https://doi.org/10.18653/v1/n18-1202

REDMON, J., FARHADI, A., & AP, C. (2018). YOLOv3. *Nutrition Reviews*.

SAHAY, R. (2020). Cosmos DB. In *Microsoft Azure Architect Technologies Study Companion*. DOI : https://doi.org/10.1007/978-1-4842-6200-9_20
SMITH, R. (2007). Tesseract OCR Engine. *Lecture. Google Code. Google Inc*.
SUN, C., QIU, X., XU, Y., & HUANG, X. (2019). How to Fine-Tune BERT for Text Classification? *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. DOI : https://doi.org/10.1007/978-3-030-32381-3_16
VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, Ł., & POLOSUKHIN, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*. DOI : https://doi.org/10.48550/arXiv.1706.03762
WANG, L. (2016). *pdf2htmlEX*. https://github.com/coolwanglu/pdf2htmlEX
YU, A. W., DOHAN, D., LUONG, M. T., ZHAO, R., CHEN, K., NOROUZI, M., & LE, Q. V. (2018). QaNet: Combining local convolution with global self-attention for reading comprehension. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*. DOI : https://doi.org/10.48550/arXiv.1804.09541

Comparaison de méthodes de prétraitement pour l'alignement de mots dans des corpus parallèles alsacien-français

Delphine Bernhard[1]
(1) Université de Strasbourg, LiLPa UR 1339, F-67000 Strasbourg
dbernhard@unistra.fr

RÉSUMÉ

L'analyse des corpus de textes dans les dialectes alsaciens est rendue difficile par la grande variation qui se rencontre à l'écrit, en l'absence d'une norme orthographique stable. Dans cet article, nous décrivons des expériences visant à améliorer la qualité de l'alignement automatique des mots français et alsaciens en prétraitant les textes. Ces prétraitements visent à réduire la variation en utilisant diverses stratégies en fonction de la langue : lemmatisation, désuffixation, clés métaphone, normalisation et réduction des tokens à leur préfixe.

ABSTRACT

Comparison of pre-processing methods for word alignment in parallel Alsatian-French corpora.

The analysis of text corpora in Alsatian dialects is made difficult by the large variation that occurs in writing, in the absence of a stable orthographic standard. In this article, we describe experiments to improve the quality of automatic alignment of French and Alsatian words by preprocessing the texts. These preprocessing operations aim to reduce variation by using various strategies depending on the language: lemmatisation, stemming, metaphone keys, normalisation and reducing tokens to their prefix.

MOTS-CLÉS : dialectes alsaciens, variation, alignement de mots.

KEYWORDS: Alsatian dialects, variation, word alignment.

1 Introduction

Les dialectes alsaciens se caractérisent par une scripturalisation non normée selon des standards orthographiques. Bien que des standards aient été proposés, ils ne sont pas largement diffusés. Cette absence de norme complique la manipulation de corpus de textes pour les dialectes alsaciens pour des applications telles que la recherche dans les corpus, l'analyse thématique, ou l'entraînement et l'application d'outils de traitement automatique des langues. Il est donc nécessaire de pouvoir identifier les variantes orthographiques dans ces corpus pour pouvoir les exploiter au mieux. La gestion de la variation a été largement abordée pour divers types de données (textes historiques, médias sociaux, dialectes). En particulier, les approches supervisées telles que celles proposées par Barteld *et al.* (2019) pour des textes en moyen bas allemand ou Hosseini *et al.* (2020) pour la mise en correspondance de toponymes nécessitent de disposer de données d'entraînement, sous la forme de paires de variantes vraies ou fausses.

Nous proposons d'exploiter des textes parallèles pour collecter des données permettant d'entraîner

Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain *(LIFT) et* Traitement Automatique des Langues *(TAL),*
Marseille, France, 14 au 15 novembre 2022 , pages 49–54,

de tels systèmes de détection automatique de variantes : les mots graphiquement similaires qui sont alignés avec la même traduction en français peuvent être considérés comme des variantes graphiques. Dans cet article, nous décrivons plus particulièrement des expériences visant à déterminer la meilleure stratégie d'alignement de mots pour un corpus parallèle alsacien-français. Nous comparons deux outils d'alignement et appliquons diverses procédures de prétraitement au corpus pour améliorer les alignements. Nous montrons que les procédures de prétraitement visant à réduire la variation sont bénéfiques, en particulier lorsqu'elles sont appliquées aux textes alsaciens.

2 Données et méthode

2.1 Corpus parallèle et lexique d'évaluation

Notre corpus regroupe des textes parallèles de divers genres : contes, pièce de théâtre, textes administratifs, sites web, chansons, recettes. Il comporte 149 119 tokens alsaciens et 155 168 tokens français, pour un total de 12 945 segments alignés. On observe un accroissement du vocabulaire bien plus important dans le corpus alsacien que dans le corpus français (voir Figure 1), ce qui s'explique notamment par l'absence de standard orthographique en alsacien.

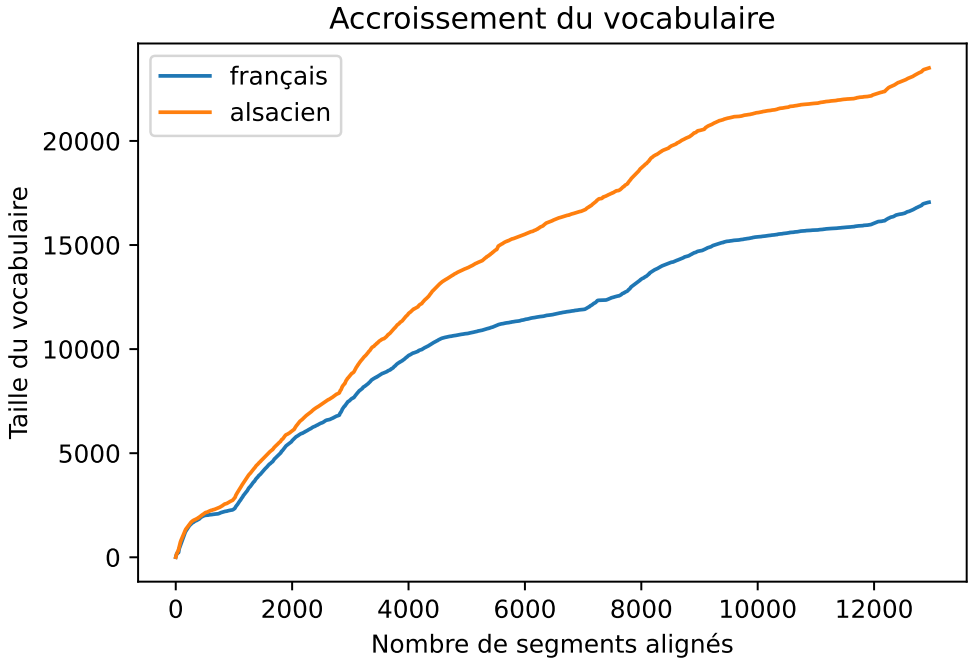


FIGURE 1 – Accroissement du vocabulaire dans les deux corpus.

La Table 1 montre quelques exemples de variantes observées dans le corpus.

français	**alsacien (variantes)**
Strasbourg	Strosburi, Strosbùri, Strossburi, Strossburig, Strossburri, Strossbùri, Stroßbùrri
bientôt	ball, boll, bàll, böl
simple	ainfàch, eifach, einfach, einfàch, ëmfàch, ëënfàch

TABLE 1 – Exemples de variantes en alsacien avec leur traduction en français.

Pour les besoins de l'évaluation, nous avons utilisé divers lexiques bilingues, et n'avons conservé que les paires de formes se trouvant dans un segment aligné de notre corpus parallèle, aboutissant ainsi à un lexique de 3 102 entrées. Nous utilisons la procédure d'évaluation proposée par Lardilleux *et al.* (2010), qui tient compte des probabilités de traduction pour le calcul de la précision du rappel et du score F1.

2.2 Outils d'alignement

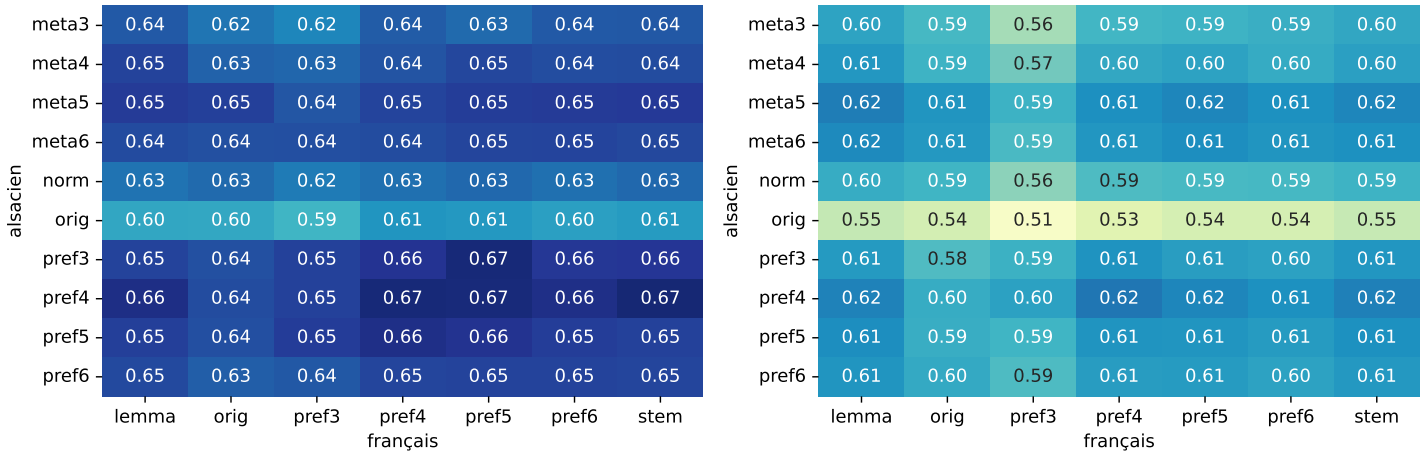
Nous avons comparé deux outils d'alignement de mots : `eflomal` (Östling & Tiedemann, 2016) [1] et `fast_align` (Dyer *et al.*, 2013) [2], avec les paramètres par défaut. Les alignements produits sont asymétriques et peuvent être rendus symétriques par diverses heuristiques. Nous utilisons ⅓ du lexique de référence pour choisir la meilleure heuristique de symmétrisation, à savoir l'intersection. Ceci est cohérent avec les travaux de Steingrímsson *et al.* (2021) qui ont également montré que l'intersection était la meilleure stratégie à la fois pour `eflomal` et `fast_align`.

2.3 Prétraitements

Nous appliquons diverses méthodes de prétraitement au corpus pour réduire la variation et ainsi améliorer la qualité de l'alignement, par rapport aux formes originales dans le corpus (`orig`) :

— français : lemmatisation (`lemma`), désuffixation (`stem`) [3] et tokens réduits à leurs `n` premiers caractères (`pref`).

— alsacien : normalisation par mise en minuscules, suppression des apostrophes, des diacritiques et remplacement des lettres doublées par un seul exemplaire (`norm`), tokens normalisés réduits à leur `n` premiers caractères (`pref`), tokens normalisés remplacés par leur clé métaphone (Philips, 2000; Bernhard, 2014) s'ils ont au moins `m` caractères (`meta`).



(a) `eflomal` (b) `fast_align`

FIGURE 2 – Scores F1 obtenus par `eflomal` et `fast_align`.

1. Version 1.0.0 publiée le 7 avril 2020 sur https://github.com/robertostling/eflomal/releases/tag/v1.0.0

2. https://github.com/clab/fast_align, Git hash `cab1e9aac8d3bb02ff5ae58218d8d225a039fa11`

3. La lemmatisation est effectuée à l'aide de spaCy 2.3.5, modèle `fr_core_news_sm` version 2.3.0. La désuffixation est obtenue à l'aide de `nltk.stem.snowball.FrenchStemmer` (`nltk` version 3.7).

3 Résultats

Les scores F1 obtenus sur les ⅔ restant du lexique d'évaluation sont présentés dans la Figure 2 ; la Figure 3 détaille la précision et le rappel pour `eflomal`. D'une manière générale, `eflomal` obtient de meilleurs résultats que `fast_align`, ce qui confirme les résultats de Steingrímsson *et al.* (2021) qui ont montré qu'`eflomal` était plus performant que `fast_align` pour les corpus de petite taille.

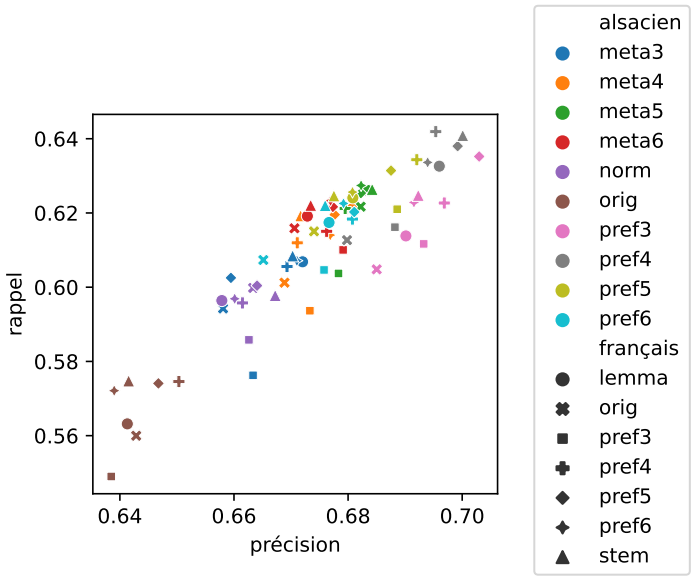


FIGURE 3 – Précision et rappel obtenus par `eflomal` pour les différents prétraitements.

Nous mesurons la significativité statistique des résultats obtenus par `eflomal` à l'aide d'une méthode de ré-échantillonnage bootstrap apparié (« paired bootstrap resampling », Berg-Kirkpatrick *et al.* (2012)) avec 1 000 réplications pour divers types de prétraitements (voir Tables 2 et 3).

Prétraitement alsacien		**Prétraitement français**						
1	**2**	lemma	orig	pref3	pref4	pref5	pref6	stem
meta5	orig	0.000	0.000	0.000	0.000	0.000	0.000	0.000
pref3	orig	0.000	0.000	0.000	0.000	0.000	0.000	0.000
pref4	orig	0.000	0.000	0.000	0.000	0.000	0.000	0.000
pref3	meta5	0.716	0.900	0.041	0.105	0.010	0.390	0.324
pref4	meta5	0.041	0.154	0.021	0.001	0.005	0.056	0.003

TABLE 2 – Valeurs de p pour les différences entre prétraitements sélectionnés pour le corpus alsacien. Les valeurs correspondant aux seuils $p < 0.05$ et $p < 0.01$ sont mises en évidence.

Prétraitement fr		**Prétraitement alsacien**									
1	**2**	meta3	meta4	meta5	meta6	norm	orig	pref3	pref4	pref5	pref6
lemma	orig	0.008	0.002	0.268	0.280	0.234	0.597	0.065	0.000	0.053	0.016
stem	orig	0.010	0.021	0.251	0.193	0.427	0.917	0.007	0.000	0.101	0.008
pref4	orig	0.020	0.108	0.406	0.347	0.302	0.021	0.005	0.000	0.000	0.007
pref4	stem	0.339	0.184	0.153	0.702	0.267	0.266	0.415	0.379	0.007	0.517
pref5	stem	0.055	0.289	0.372	0.375	0.496	0.353	0.024	0.316	0.042	0.391
pref4	lemma	0.366	0.030	0.195	0.576	0.413	0.052	0.064	0.825	0.010	0.315
pref5	lemma	0.064	0.321	0.413	0.249	0.151	0.057	0.002	0.184	0.086	0.262

TABLE 3 – Valeurs de p pour les différences entre prétraitements sélectionnés pour le corpus français. Les valeurs correspondant aux seuils $p < 0.05$ et $p < 0.01$ sont mises en évidence.

L'application de prétraitements au corpus alsacien (métaphone, préfixe) conduit systématiquement à de meilleurs résultats par rapport au corpus original. L'étude de significativité (Table 2) montre que la différence observée entre les prétraitements `meta5`, `pref3` et `pref4` et le corpus d'origine `orig`, non prétraité, est statistiquement significative ($p < 0.05$). Il est toutefois plus difficile de conclure sur la supériorité du prétaitement par préfixe (`pref3` et `pref4`) par rapport au prétraitement par métaphone (`meta5`) : pour ces prétraitements, la différence n'est statistiquement significative que dans quelques cas, avec certains types de prétraitements pour le français. Globalement, nos résultats pour l'alsacien vont dans le sens des observations faites par Östling & Tiedemann (2016) qui montrent qu'une méthode de désuffixation approximative consistant à ne conserver que les 4 premières lettres d'un mot permet généralement d'améliorer les résultats en traduction automatique statistique pour les langues à suffixation.

Par ailleurs, si l'on observe des résultats légèrement meilleurs avec un prétraitement pour le français, les différences ne sont généralement pas statistiquement significatives (Table 3), en particulier lorsque l'on compare les trois méthodes de prétraitement proposées : préfixes, lemmatisation et désuffixation.

Enfin, la Table 4 donne quelques exemples de mots alsaciens avec leur traduction la plus probable en l'absence de prétraitement des corpus et avec réduction à un préfixe de 4 caractères. La traduction attendue est également indiquée.

alsacien	**traduction attendue**	**sans prétraitement**	**préfixation à 4**
Biehn	grenier	∅	grenier
Dràche	dragon	dragon	dragon
fàrwig	coloré	original	coloré
schloeje	battre	jusqu'	puis
schwàsiere	choisir	Nouvel, désirés	∅
versteckle	cacher	cacher	cacher, Cacher
wascha	laver	∅	laver

TABLE 4 – Exemples de mots alsaciens avec leur traduction la plus probable.

4 Conclusion et perspectives

Nous avons présenté des expériences visant à comparer deux outils d'alignement de mots en appliquant diverses procédures de prétraitement à des corpus parallèles alsacien-français. Les résultats montrent une supériorité d'`eflomal` sur `fast_align`, ainsi que la pertinence d'un prétraitement simple consistant à normaliser les mots alsaciens et à les réduire à leurs préfixes, les meilleurs résultats étant obtenus pour un préfixe de 4 lettres. L'utilisation des clés métaphone permet aussi d'obtenir de bons résultats. L'avantage du prétraitement est moins évident pour le français, même si les divers types de prétraitements comparés (préfixe, lemmatisation, désuffixation) peuvent conduire à des améliorations en combinaison avec certains prétraitements pour l'alsacien.

Dans la suite de ce travail, nous souhaitons améliorer la prise en compte de la variation graphique en nous inspirant de la méthode proposée par Burlot & Yvon (2017) pour l'alignement entre langues morphologiquement complexes et langues plus simples. Les lexiques bilingues nous servirons ensuite de données d'entraînement pour la détection automatique de variantes en alsacien.

Remerciements

Ces travaux ont été réalisés dans le cadre du projet ANR-21-CE27-0004 DIVITAL soutenu par l'Agence Nationale de la Recherche.

Nous remercions les étudiantes ayant participé à la collecte et à l'alignement des phrases du corpus parallèle : Natália Leščišínová, Camille Meyer et Johana Libman.

Références

BARTELD F., BIEMANN C. & ZINSMEISTER H. (2019). Token-based spelling variant detection in Middle Low German texts. *Language Resources and Evaluation*, p. 1–30. DOI : 10.1007/s10579-018-09441-5.

BERG-KIRKPATRICK T., BURKETT D. & KLEIN D. (2012). An Empirical Investigation of Statistical Significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, p. 995–1005, Jeju Island, Korea.

BERNHARD D. (2014). Adding Dialectal Lexicalisations to Linked Open Data Resources : the Example of Alsatian. In *Proceedings of the Workshop on Collaboration and Computing for Under Resourced Languages in the Linked Open Data Era (CCURL 2014)*, p. 23–29, Reykjavík, Iceland. HAL : hal-00966820.

BURLOT F. & YVON F. (2017). Learning Morphological Normalization for Translation from and into Morphologically Rich Languages. *The Prague Bulletin of Mathematical Linguistics*, **108**(1), 49–60. DOI : 10.1515/pralin-2017-0008.

DYER C., CHAHUNEAU V. & SMITH N. A. (2013). A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 644–648, Atlanta, Georgia : Association for Computational Linguistics.

HOSSEINI K., NANNI F. & COLL ARDANUY M. (2020). DeezyMatch : A Flexible Deep Learning Approach to Fuzzy String Matching. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 62–69, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-demos.9.

LARDILLEUX A., GOSME J. & LEPAGE Y. (2010). Bilingual lexicon induction : Effortless evaluation of word alignment tools and production of resources for improbable language pairs. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, p. 252–256. HAL : hal-00488768.

PHILIPS L. (2000). The Double Metaphone Search Algorithm. *C/C++ Users Journal*, **18**(6), 38–43.

STEINGRÍMSSON S., LOFTSSON H. & WAY A. (2021). CombAlign : a Tool for Obtaining High-Quality Word Alignments. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, p. 64–73, Reykjavik, Iceland (Online) : Linköping University Electronic Press, Sweden.

ÖSTLING R. & TIEDEMANN J. (2016). Efficient Word Alignment with Markov Chain Monte Carlo. *The Prague Bulletin of Mathematical Linguistics*, **106**(1), 125–146. DOI : 10.1515/pralin-2016-0013.

Compositionality in a simple corpus

Manuel Vargas Guzmán[1,2], Maria Boritchev[1], Jakub Szymanik[3], Maciej Malicki[1]
(1) IMPAN, Jana i Jędrzeja Śniadeckich 8, 00-656 Warsaw, Poland
(2) University of Warsaw, Krakowskie Przedmieście 26/28, 00-927 Warsaw, Poland
(3) CIMeC and DISI - University of Trento Rovereto (TN) - Italy
m.vargas-guzman@uw.edu.pl, mboritchev@impan.pl, mmalicki@impan.pl,
jakub.szymanik@gmail.com

RÉSUMÉ

Nous étudions la capacité des réseaux de neurones à apprendre des structures compositionnelles en nous concentrant sur un corpus logique simple bien défini, et sur les phénomènes de compositionnalité centrés sur les preuves. Nous menons notre étude dans un cadre minimal en créant un corpus logique simple, où tous les phénomènes liés à la compositionnalité proviennent de la structure des preuves, car toutes les phrases du corpus sont des implications en logique propositionnelle. En entraînant des réseaux de neurones sur ce corpus, nous testons différents aspects de la compositionnalité, à travers des variations de la longueur des preuves et des permutations des constantes en jeu.

ABSTRACT

We investigate the capacity of neural networks (NNs) to learn compositional structures by focusing on a well-defined simple logical corpus, and on proof-centered compositionality. We conduct our investigation in a minimal setting by creating a simple logical corpus, where all compositionality-related phenomena come from the structure of proofs as all the sentences of the corpus are propositional logic implications. By training NNs on this corpus we test different aspects of compositionality, through variations of proof lengths and permutations of the constants.

MOTS-CLÉS : Compositionalité, logique, raisonnement, réseaux de neurones.

KEYWORDS: Compositionality, logic, reasoning, neural networks.

1 Introduction

Compositionality is a vastly discussed subject across natural language semantics, logic, but also natural language processing and nowadays, neural networks. Partee (1984) defines compositionality as the principle according to which"the meaning of an expression is a function of the meanings of its parts and of the way they are syntactically combined". In the past years, the investigation of the capacity of neural networks (NNs) to compositionally use/produce/deduce rules and/or sentences has gained more and more importance in the natural language processing field, in particular through the scope of natural language understanding tasks. Oscillating between mathematics and sentences in English, works such as Bowman *et al.* (2015), Saxton *et al.* (2019), and most recently Ontanon *et al.* (2022) show different ways in which NNs can be seen as more or less compositional, depending on

Research supported by the National Center of Science [grant 2020/37/B/HS1/04220]

the task and the mean of testing.

Bowman *et al.* (2015) test the capacity of LSTMs, using explicit clues, to discover and implicitely use recursive compositional structures. To do so, the authors implement the task as a classification one, outputting one of 7 possible logical relations between a given pair of sentences. Saxton *et al.* (2019) take a very different approach to the compositionality question by investigating the capacity of neural models to solve mathematical problems, in English, of various types (arithmetic, algebra, probability, calculus). The goal of the authors is to study the capacity of Transformers and Recurrent Neural Networks (RNNs) to compose and generalize mathematical concepts and operations. The results and performances vary greatly from one type of mathematical problem to another. Ontanon *et al.* (2022) introduce a dataset designed to evaluate the capacity of NNs to perform logical inferences. The dataset is composed of sentences that use propositional logic, a fragment of first order logic, and English. Hupkes *et al.* (2020) present a systematic set of tests for NNs capacity to compositionally generalize a rule set: (1) capacity of the NN to recombine known parts and rules to produce results it has never been exposed to before; (2) capacity of the NN to extend its predictions to data longer than the one it has been exposed to in training; (3) preference of the NN to compose in a local or in a global way; (4) robustness of the NN's predictions w.r.t synonym substitution; (5) preference of the NN towards rules or exceptions during training. The authors then instantiate the test suite on an artificial dataset and apply it to a RNN, a convolution-based neural network, and a transformer. The work presented in our article is placed in the footsteps of this latter approach.

The motivation for our work is two-fold: we want to investigate the capacity of neural networks to produce natural reasoning, but the approach we are taking grows from mathematical reasoning first. Indeed, implication is a fundamental element of natural language inference and understanding, as well as logic. Because of this, we begin by considering a simple corpus of propositional logic implications. Following approaches developed in previous work, in particular Hupkes *et al.* (2020), we focus on compositionality. To our knowledge, we are the first to study inference in the presence of multiple premises, and to work specifically on proof-structure compositionality and its different aspects. We do a fine-grained analysis of the output errors of our models by computing the Hamming distances between expected outputs and actual outputs of our models. As Hamming distances measure the number of differences between the two compared vectors, they constitute a tool that allows us to quantify how far away from the right answer the wrong answers are. In the following, we present our data in section 2, then we develop our experimental set-up in section 3. In this section we conduct an error analysis, and we introduce two compostionality tests. The data, code and materials for this article are shared in the GitHub respository `https://github.com/MBoritchev/compostitionality-simple-corpus`.

2 A simple corpus

The logical language of our corpus is defined as follows: let $\mathcal{C} = \{X_1 \dots X_n\}$ be a set of *constants*, we build *formulas* as logical implications: "$X_i \rightarrow X_j$". Then, we define a *knowledge base* $\mathcal{KB}$ which is a set of formulas, called *premises*. From a $\mathcal{KB}$, we can prove new formulas by using a unique derivation rule:

$$\frac{X_i \rightarrow X_j \qquad X_j \rightarrow X_k}{X_i \rightarrow X_k}$$

Given a $\mathcal{KB}$, a formula h is a *valid hypothesis* or a *conclusion*, if $h \in \mathcal{KB}$ or if it can be derived from

a set of premises $\pi \subseteq \mathcal{KB}$ by using the derivation rule. We write $\pi \vdash h$ to denote that there exists such a proof; if h cannot be proved from $\mathcal{KB}$, we write $\mathcal{KB} \nvdash h$, and call it an *invalid hypothesis*.

We represent $\mathcal{KB}$ as a directed graph $G = (V, E)$ where V is a set of vertices and $E \subseteq \{(u, v) \mid (u, v) \in V^2 \text{ and } u \neq v\}$ is a set edges. In this graph, each vertice is a constant and each edge corresponds to a premise. A proof corresponds then to a (directed) path between two vertices.

In our study, $\mathcal{KB}$ has the form of a tree. Since in this case, there is at most one path between any two vertices, for any h such that $\mathcal{KB} \vdash h$, there is a unique shortest proof, comprised of premises from a set π, that witnesses it. In particular, the *length* of the proof of h is the size of π.

Example : Let $\mathcal{KB}$ be the graph from figure 1, then $\mathcal{KB} \vdash X_3 \rightarrow X_{24}$ as there is a path from vertice X_3 to vertice X_{24} in the graph, in other words, $X_3 \rightarrow X_{24}$ can be derived from $\pi = \{X_3 \rightarrow X_9, X_9 \rightarrow X_{13}, X_{13} \rightarrow X_{24}\}$; therefore, $X_3 \rightarrow X_{24}$ is a valid hypothesis/conclusion. On the other hand, we have that $\mathcal{KB} \nvdash X_{30} \rightarrow X_{27}$ since there is no path that connects those two vertices and hence, the formula $X_{30} \rightarrow X_{27}$ is an invalid hypothesis.

We investigate neural models such that, for a given knowledge base $\mathcal{KB}$ and a hypothesis h, the model provides the necessary set π of premises from $\mathcal{KB}$ to prove h, if they exist. Otherwise, it indicates that there is no such π. We consider two architectures: a multilayer perceptron (MLP) and a recurrent neural network (RNN).

Data encoding To train a model, we use a multi-label approach that consists of a neural network $f(X) = \hat{y}$, where the input $X = [\mathcal{KB}, h]$ is a vector that encodes the set $\mathcal{KB} = \{p_1, \ldots, p_n\}$ of all premises and a hypothesis h, and the output $\hat{y}$ is the predicted value of the label y, a binary vector of size n such that for each element $i \in y$

$$i = \begin{cases} 1 & \text{if } p_i \in \pi \\ 0 & \text{otherwise} \end{cases}$$

where $\pi \subseteq \mathcal{KB}$ is the set of necessary premises to prove h. Moreover, if $\mathcal{KB} \nvdash h$, then y consists of n zeros. Therefore in the sequel we also refer to invalid hypotheses as hypotheses with proof length 0, premises in $\mathcal{KB}$ are hypotheses with proof length 1, etc.

Example : Let $\mathcal{KB} = \{X_1 \rightarrow X_3,\ X_3 \rightarrow X_6,\ X_6 \rightarrow X_4,\ X_6 \rightarrow X_2,\ X_6 \rightarrow X_5\}$ and $h = X_1 \rightarrow X_5$. Then, the vectors X and y are built as follows:

$$X = [X_1 \rightarrow X_3,\ X_3 \rightarrow X_6,\ X_6 \rightarrow X_4,\ X_6 \rightarrow X_2,\ X_6 \rightarrow X_5,\ X_1 \rightarrow X_5]$$
$$y\ = [1\ 1\ 0\ 0\ 1]$$

The above vector encodes a proof of length 3.

Each formula from input X is encoded in a vector of dimension $2n$ (where n is the the size of the set $\mathcal{C}$) as a *one-hot* fashion. For instance, let $\mathcal{C} = \{X_1,\ X_2,\ X_3,\ X_4,\ X_5\}$, then the formula $X_2 \rightarrow X_5$ is encoded as

$$[0\ 1\ 0\ 0\ 0 \mid 0\ 0\ 0\ 0\ 1]$$

where the first n digits represent X_2 and the constant X_5 is encoded within the last n bits from the vector.

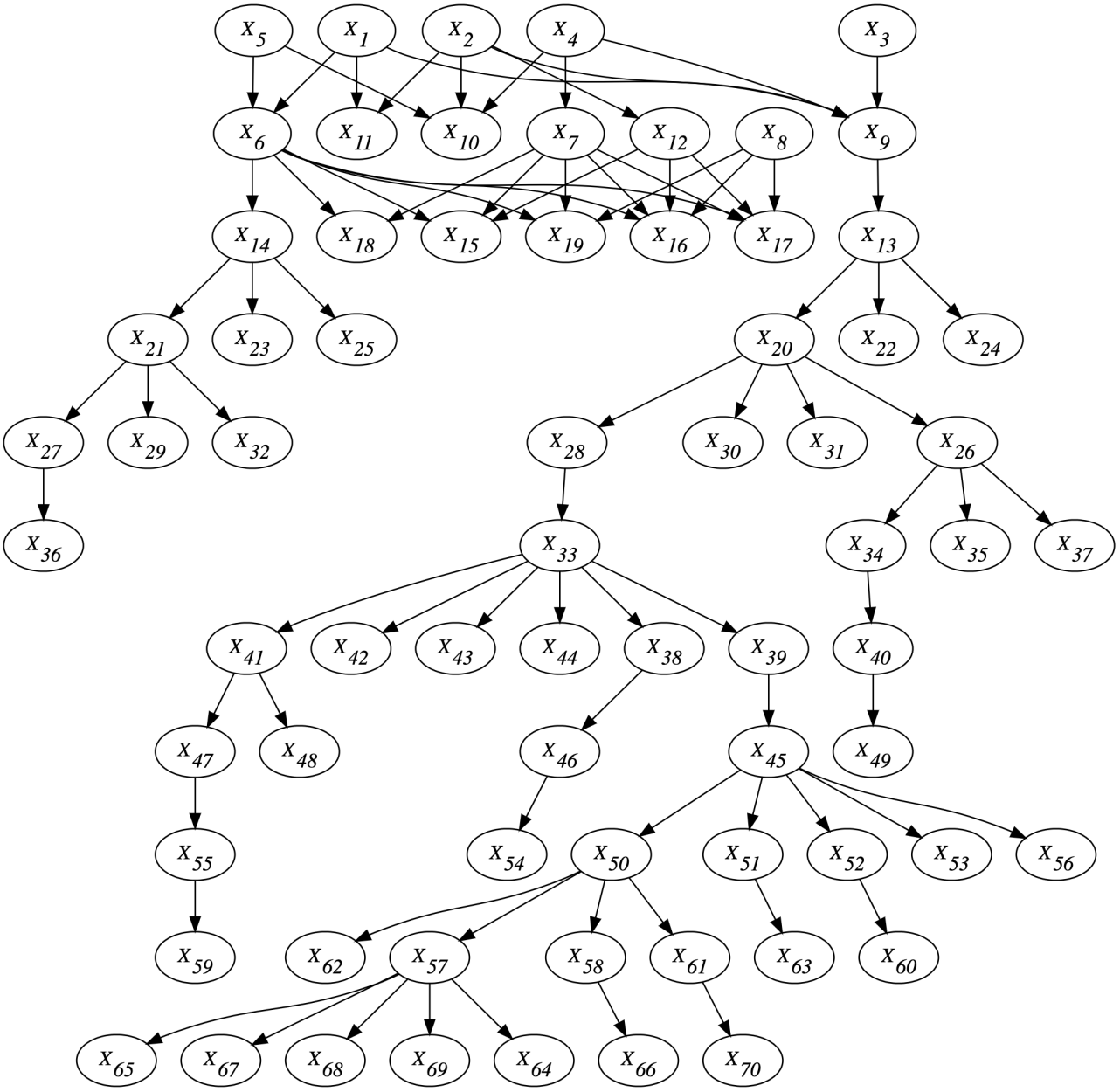


FIGURE 1 – A Knowledge Base built with 70 constants and 82 formulas

3 Tests and limits

In the preliminary study we present here, we only consider one $\mathcal{KB}$, depicted in figure 1: it is randomly built, with 70 constants and 82 premises. Among the $4{,}830 = 70 \times 69$ hypothesis, 531 are valid in the $\mathcal{KB}$.

3.1 Initial experiment

First, we trained and tested our models with 75/25% split of the $\mathcal{KB}$, stratified by length of proofs. Table 1 shows the detailed data distribution for each class.

Length of proofs	Total data	Train data	Test data
0	4299	3224	1075
1	82	62	20
2	75	56	19
3	63	47	16
4	60	45	15
5	51	38	13
6	54	40	14
7	41	31	10
8	42	32	10
9	35	26	9
10	28	21	7

TABLE 1 – Data distribution for our initial experiment

Neural networks setup We trained both the MLP and the RNN using the *Adamax* algorithm to optimize the weight values with its default learning rate of 0.001. The MLP has a single hidden layer of 2500 neurons, and the RNN is equiped with two hidden layers, each with 200 neurons. For both architectures, the *hyperbolic tangent* function (*tanh*) is used in the hidden layers, and the *sigmoid* function in the output layer. Moreover, every layer has its respective *bias* weight. We run between 200 and 300 epochs with a *batch size* of 20.

Accuracy The overall performance of the models is as follows: 97.76% and 97.93% of correct predictions for the MLP and the RNN, respectively, and for valid hypotheses, the MLP achieved 80.45% and the RNN predicted correctly 82.71% of the test data. The results by length of proofs (table 2) show a counter-intuitive shape: the models have poor accuracy for proofs of length 1, which correspond to the task of recognizing the conclusion in the set of formulas; then, as the length of the proofs goes up, so does accuracy, reaching 100% at length 5, with one exception.

This result is at least partly explained by the exploration of data we performed: the structure of our $\mathcal{KB}$ entails that for long proofs (length ≥ 2), the model sees both the long proof and a number of its sub-proofs in training. Therefore, the longer the proof, the more the model has learned about it, see table 2 for details.

We computed the Hamming distances between the expected output and the rounded up actual output of our NNs.

Example : Let y be the expected output vector (label) for our NN, $\hat{y}$ the actual output vector (prediction):

$$y = [1\ 1\ 0\ 0\ 1]$$
$$\hat{y} = [0.68\ 0.98\ 0.33\ 0.12\ 0.46]$$

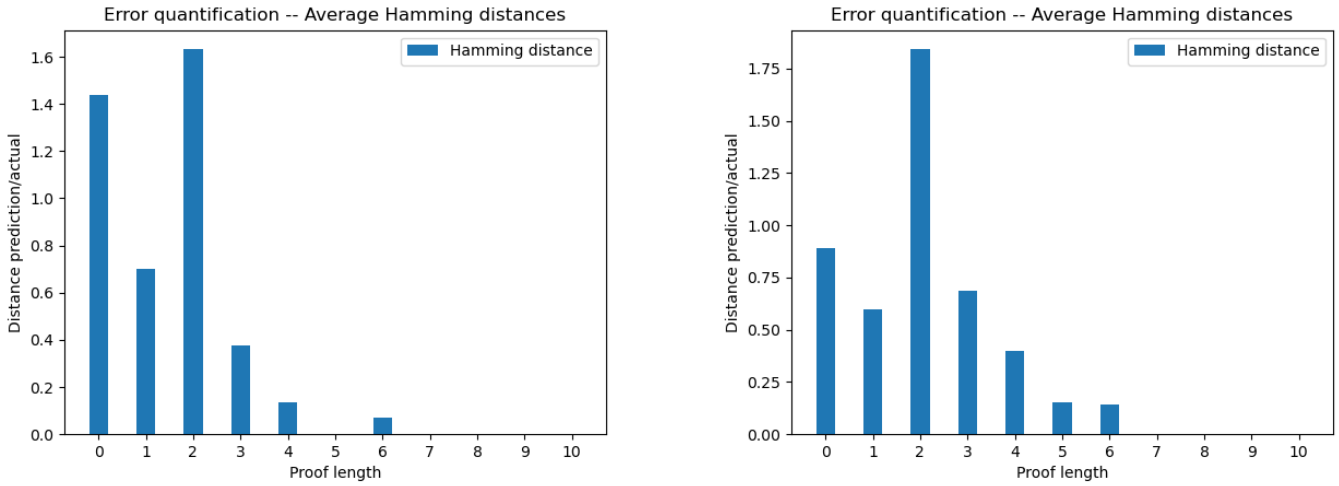
Then $\hat{y}$ rounded is $[1\ 1\ 0\ 0\ 0]$. The Hamming distance between y and $\hat{y}$ rounded is equal to 1, as the two vectors differ only in one of the coordinates, the last one.

Figure 2 shows the average Hamming distances. For proofs of length 1, these were at most 1, which corresponds to only one wrong selected/not-selected formula, which suggests that the NNs do perform the task that is expected of them while getting the formula wrong. For length 2, the distances were at

Length of proofs	MLP test	RNN test	hypothesis	# of sub-proofs	# of formulas
0	99.91%	99.81%	3224	0	0
1	40.00%	50.00%	62	62	62
2	57.89%	63.16%	56	138	194
3	68.75%	75.00%	47	202	362
4	93.33%	93.33%	45	331	695
5	100.00%	100.00%	38	416	1017
6	100.00%	100.00%	40	640	1756
7	100.00%	100.00%	31	645	2002
8	100.00%	90.00%	32	839	2869
9	100.00%	100.00%	26	845	3178
10	100.00%	100.00%	21	827	3384

TABLE 2 – Models achieve better accuracy on longer proofs because of overlaps in training

most 2, and for all larger lengths, the average distances are lower than 0.5, for MLP and RNN alike.



(a) MLP

(b) RNN

FIGURE 2 – Average Hamming distances between the expected output and the actual output for the initial experiment

3.2 Compositionality tests

Then, we explored compositionality tests in the context of our simple corpus, through: (1) variations in the number of formulas needed to prove a hypothesis; (2) permutations in the order of constants. (1) allows us to test the productivity of implication: as implication is transitive, it can be composed; then, with (2), we want to test the capacity of the model to abstract away from the order of constants, which is irrelevant for the derivation rule considered in the study.

For (1), the NNs are trained on proofs from length n_1 to n_2, $n_1 < n_2$; then, we test their performance by predicting all unseen hypothesis from the $\mathcal{KB}$. Test (1) is split in two parts. First, we train the NNs on proofs of length 0 to n_2 and then test them on proofs of length larger than n_2; this is the *unseen*

longer proofs setting. Second, we train the NNs on proofs of length n_1 to 10 and then test them on proofs of length smaller than n_1; this is the *unseen shorter proofs* setting.

Accuracies Table 3 shows the results of test (1) in two settings: 3a when the difference between train and test are the longer proofs, 3b when those are the shorter proofs. For the unseen longer proofs test, we obtained better results by changing the train/test data split only for invalid hypothesis to 20/80%. The models perform better predicting unseen longer proofs than unseen shorter proofs, though in both cases the accuracy goes down quickly. The RNN shows a better accuracy than the MLP.

Train data	MLP test	RNN test
0-9	100.0%	100.0%
0-8	92.06%	93.65%
0-7	73.33%	82.86%
0-6	45.89%	65.07%
0-5	17.0%	46.0%

(a) Unseen longer proofs

Train data	MLP test	RNN test
1-10	44.1%	48.08%
2-10	40.95%	41.61%
3-10	37.5%	40.13%
4-10	7.37%	11.06%
5-10	1.81%	2.16%

(b) Unseen shorter proofs

TABLE 3 – Compostionality tests: variations in the number of formulas

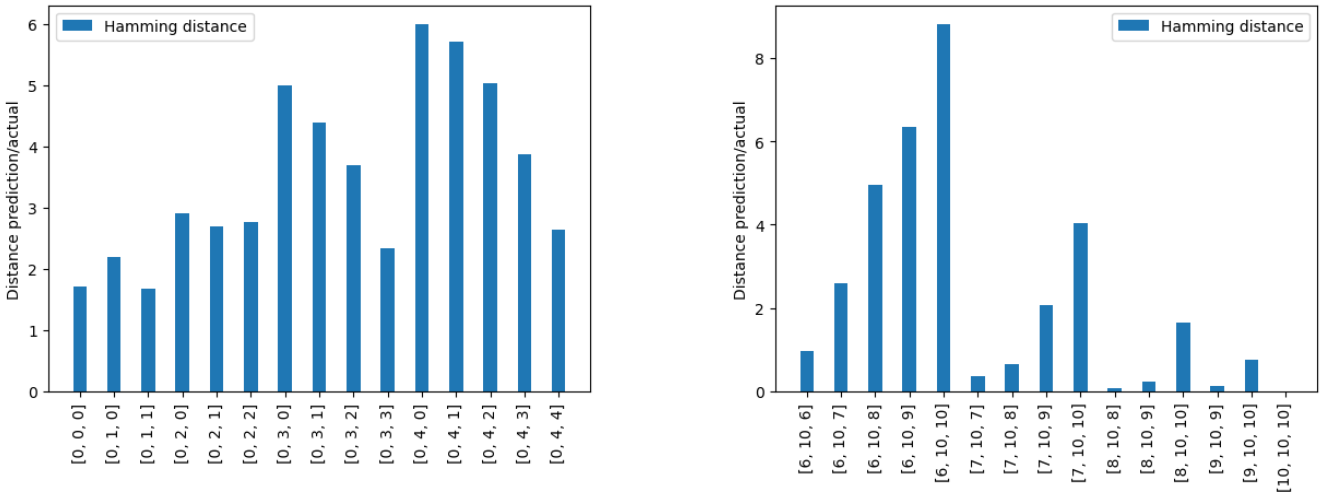
Test (2) is the following: if the model has been trained on a given $\mathcal{KB}$, then a permutation function $\sigma : \mathcal{C} \to \mathcal{C}$ has been applied to $\mathcal{KB}$ to obtain $\mathcal{KB}$', how will the model behave with inputs from $\mathcal{KB}$'? Our first results are straightforward, the models are incapable of adapting to permutation of constants, the accuracy is of 17.84% for MLP, 66.43% for RNN for invalid hypotheses, and 0% for both models for all valid hypotheses, with the exception of a 1.33% for RNN for length 2.

Figures 3 and 4 show the average Hamming distances for experiment (1). In each triple [n1,n2,m] on the x-axis, n1 and n2 stand for the minimal and maximal lengths of unseen proofs, and m corresponds to the length of evidence on which the model is being tested: for example, the triple [0,3,1] corresponds to the model that has not been exposed to proofs of length 0 to 3 in training and is being tested on proofs of length 1. The shape of the curves is similar to the ones for the initial experiment, while the values of the average Hamming distances are quite large. In particular, for unseen lengths 6-10 for tested length 10, where the distance is higher than 8 for MLP, higher than 3.5 for RNN. Thus, the analysis of Hamming distances corroborates the results of the initial experiment.

4 Conclusion

Even though NNs appear to be able to pick up some structure from data, our results show that the models have a hard time generalizing it to unseen proof lengths. Moreover, NNs appear to be substantially more sensitive to the order of constants than to the overall structure of the $\mathcal{KB}$. On the other hand, the experiments show an increase in performance for longer seen lengths of proofs, which, correlated with the number of sub-proofs seen by the model in training, suggests some amount of compositionality. Our hypothesis is that the NNs may be able to use information learned on smaller lengths of proofs to improve the performance for larger lengths.

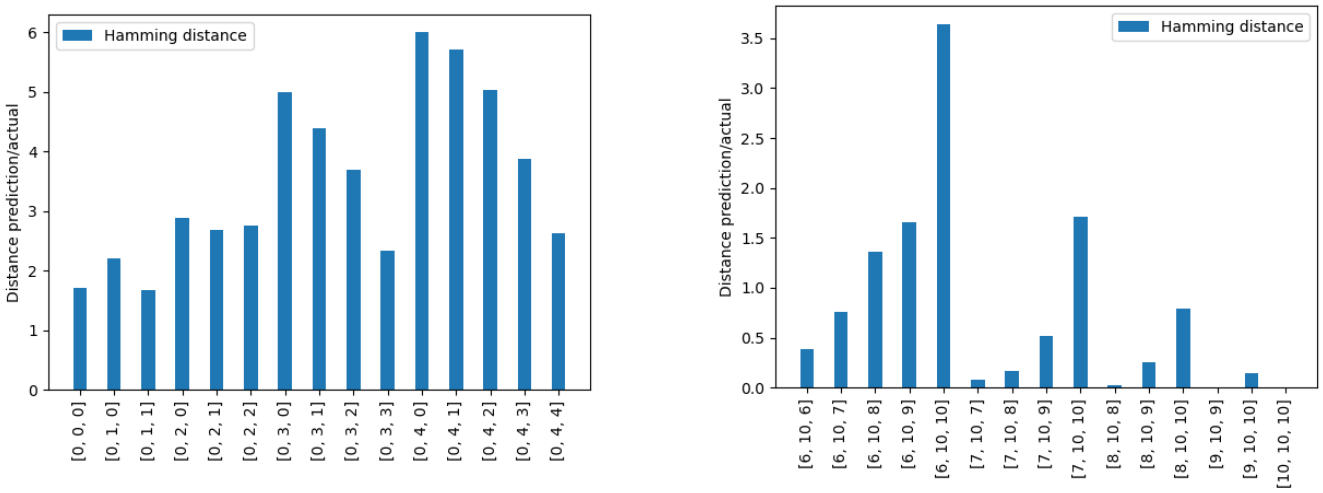
The compositionality tests that we performed show results that are consistent with ones presented in



(a) Minimal length of unseen proof : 0

(b) Maximal length of unseen proof : 10

FIGURE 3 – Average Hamming distances between the expected output and the actual output for the unseen length experiment, MLP



(a) Minimal length of unseen proof : 0

(b) Maximal length of unseen proof : 10

FIGURE 4 – Average Hamming distances between the expected output and the actual output for the unseen length experiment, RNN

Hupkes *et al.* (2018). When confronted with sequences longer than the ones they where trained on, the accuracy of NNs from Hupkes *et al.* (2018) drops significantly.

The preliminary study we present here shows that the overall performance cannot be used as the main tool in evaluating the capacity of a model to compositionally select the right premises to prove a given conclusion. The question this observation rises is what is the representation of the data that the NN builds in training? We would like to investigate this through visualisation and diagnostic techniques for NNs such as the ones presented in Hupkes *et al.* (2018).

An interesting remark that has been raised to us is the fact that our dataset is relatively small. Another direction of investigation we are undertaking has to do with this dataset size : for economi-

nal/ecological/ethical reasons, we would like to run our training experiments on the smallest possible datasets while not compromising on the quality of results. Therefore, we are conducting a comparative study through different dataset sizes. Following the same logic, we also investigate other types of encoding. The next step for our research will be its extension to other architectures of NNs, such as Transformers.

Références

BOWMAN S. R., MANNING C. D. & POTTS C. (2015). Tree-structured composition in neural networks without tree-structured architectures. *arXiv preprint arXiv :1506.04834*.

HUPKES D., DANKERS V., MUL M. & BRUNI E. (2020). Compositionality decomposed : how do neural networks generalise ? *Journal of Artificial Intelligence Research*, **67**, 757–795.

HUPKES D., VELDHOEN S. & ZUIDEMA W. (2018). Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, **61**, 907–926.

ONTANON S., AINSLIE J., CVICEK V. & FISHER Z. (2022). Logicinference : A new datasaet for teaching logical inference to seq2seq models. In *ICLR2022 Workshop on the Elements of Reasoning : Objects, Structure and Causality*.

PARTEE B. (1984). Compositionality. *Varieties of Formal Semantics*, **3**, 281–311.

SAXTON D., GREFENSTETTE E., HILL F. & KOHLI P. (2019). Analysing mathematical reasoning abilities of neural models. *arXiv preprint arXiv :1904.01557*.

Dependency Parsing with Backtracking using Deep Reinforcement Learning

Franck Dary, Maxime Petit, Alexis Nasr
Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
{franck.dary,maxime.petit,alexis.nasr}@lis-lab.fr

RÉSUMÉ

Les algorithmes gloutons pour le TAL, tels que l'analyse syntaxique en tranisition souffrent du problème de la propagation d'erreurs. Une manière de lutter contre ce problème consiste à effectuer des retours arrière afin d'explorer des solutions alternatives lorsque la solution suivie mène à une impasse. Pour mettre en œuvre un tel comportement, on utilise l'apprentissage par renforcement qui permet à l'algorithme d'apprendre à effectuer un retour arrière lorsque ce dernier lui procure une récompense supérieure à celle reçue en continuant à explorer la solution courante. Nous montrons dans ce travail qu'une telle methode peut être mise en œuvre avec succès pour l'analyse en transition et l'étiquetage en parties de discours.

ABSTRACT

Here the title in English.

Greedy algorithms for NLP such as transition based parsing are prone to error propagation. One way to overcome this problem is to allow the algorithm to backtrack and explore an alternative solution in cases where new evidence contradicts the solution explored so far. In order to implement such a behavior, we use reinforcement learning and let the algorithm backtrack in cases where such an action gets a better reward than continuing to explore the current solution. We test this idea on both POS tagging and dependency parsing and show that backtracking is an effective means to fight against error propagation.

MOTS-CLÉS : Analyse syntaxique en transitions, apprentissage par renforcement, retour arrière.

KEYWORDS: Transition based parsing, reinforcement learning, backtracking.

Transition based parsing has become a major approach in dependency parsing, since the work of Yamada & Matsumoto (2003) and Nivre *et al.* (2004) for it combines linear time complexity and high linguistic performances. The algorithm follows a local and greedy approach to parsing that consists in selecting at every step of the parsing process the action that maximizes a local score, typically computed by a classifier. The action selected is greedily applied to the current configuration of the parser and yields a new configuration.

At training time, an oracle function transforms the correct syntactic tree of a sentence into a sequence of correct (configuration, action) pairs. These pairs are used to train the classifier of the parser. The configurations that do not pertain to the set of correct configurations are never seen during training.

At inference time, if the parser predicts and executes an incorrect action, it produces an incorrect configuration, with respect to the sentence being parsed, which might have never been seen during training, yielding a poor prediction of the next action to perform. Besides, the parser follows a single

Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain *(LIFT)* et Traitement Automatique des Langues *(TAL)*,
Marseille, France, 14 au 15 novembre 2022 , pages 64–66,

hypothesis by greedily selecting the best scoring action. The solution built by the parser can be sub-optimal for there is no guarantee that the sum of the scores of the actions selected maximizes the global score.

These are well known problems of transition based parsing and several solutions have been proposed in the literature to overcome them. The solution we propose in this work consists in allowing the parser to backtrack. At every step of the parsing process, the parser has the opportunity to undo its n previous actions to explore alternative solutions. The decision to backtrack or not is taken each time a new word is considered, before trying to process it, by giving the current configuration to a binary classifier, that will assign a score to the backtracking action. Traditional supervised learning is not suited to learn such a score, since the training data contains no occurrences of backtrack actions. In order to learn in which situation a backtrack action is worthy, we use reinforcement learning.

This work is based on a formal machine, called a Backtracking Reading machine and one form of reinforcement learning called deep Q learning. Both are briefly described below.

Backtracking Reading Machines

Our model is an extension of the Reading Machine, a general model for NLP proposed in Dary & Nasr (2021) that generalizes transition based parsing to other NLP tasks. A Reading Machine is a finite automaton which states correspond to linguistic levels. There can be, for example, one state for POS tagging, one state for lemmatization, one state for syntactic parsing...When the machine is in a given state, an *action* is predicted, which generally writes on an output tape a label corresponding to the prediction just made.

We augment the machines described above in order to allow them to undo some preceding actions. This ability relies on three elements : (a) the definition of a new action, called BACK, that undoes a certain number of actions (b) the history of the actions performed so far in order to decide which actions to undo and (c) the definition of undoing an action.

Deep reinforcement learning

Reading Machines, as introduced by Dary & Nasr (2021) are trained in a supervised learning fashion. Given data annotated at several linguistic levels, an oracle function decomposes it into a sequence of configurations and actions $(c_0, a_0, c_1, a_1, \ldots, c_n, a_n)$. This sequence of configurations and actions constitute the training data of the classifiers of the machine to train : pairs (c_i, a_i) are presented iteratively to the classifier during the training stage. A backtracking Reading Machine cannot be trained this way since there are no occurrences of BACK actions in the data. In order to learn useful occurrences of such actions, the training process should have the ability to generate some BACK actions and be able to measure if this action was beneficial. In order to implement such a behavior, we use Reinforcement Learning, more specifically a simple form of deep Q learning (Mnih *et al.*, 2013) and approximate function Q, which is at the heart of the model, using a multi-layered perceptron.

Experiments and results

We experiment our ideas on seven languages, on two tasks : POS tagging and syntactic parsing, seen as independent tasks or trained jointly and show that backtraking yields better results in all configurations.

Références

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

DARY F. & NASR A. (2021). The reading machine : A versatile framework for studying incremental parsing strategies. In *Proceedings of the 17th International Conference on Parsing Technologies (IWPT 2021)*, p. 26–37, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.iwpt-1.3.

MNIH V., KAVUKCUOGLU K., SILVER D., GRAVES A., ANTONOGLOU I., WIERSTRA D. & RIEDMILLER M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv :1312.5602*.

NIVRE J., HALL J. & NILSSON J. (2004). Memory-based dependency parsing. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, p. 49–56.

YAMADA H. & MATSUMOTO Y. (2003). Statistical dependency analysis with support vector machines. In *Proceedings of the eighth international conference on parsing technologies*, p. 195–206.

Déterminer la similarité entre deux langues à l'aide des modèles pré-entraînés de la parole. Une étude pilote

Séverine Guillaume Guillaume Wisniewski
Langues et Civilisations à Tradition Orale (LACITO), UMR 7107 CNRS - Sorbonne Nouvelle - INALCO
Laboratoire de Linguistique Formelle (LLF), UMR 7110 CNRS - Université de Paris Cité
severine.guillaume@cnrs.fr guillaume.wisniewski@u-paris.fr

Résumé

Motivés par les résultats obtenus avec des modèles neuronaux comme `wav2vec` pour la transcription de langues « rares » et/ou « peu dotées », nous souhaitons pousser ces modèles encore plus loin en montrant qu'ils peuvent apporter un autre type d'aide aux linguistes dans leur travail de documentation et d'anlayse des langues en permettant d'extraire automatiquement des informations typologiques (inventaire de phonèmes, indices de complexité phonologique et morphosyntaxique, ...), d'enregistrements audio. Dans cette étude pilote, nous décrivons une première série d'expériences visant à déterminer dans quelle mesure les modèles de la parole multilingue peuvent être utilisés pour détecter des langues « similaires » au plan phonético-phonologique.

Abstract

Probing `wav2vec` for Typological Signal. A Pilot Study

Motivated by the results obtained with neural models such as `wav2vec` for the transcription of "rare", "under-resourced" languages, we wish to push these models even further by showing that they can bring another kind of help to linguists documenting and anlayzing languages by allowing the automatic extraction of typological information (phoneme inventory, phonological and morphosyntactic complexity phonological and morphosyntactic complexity, ...) from audio recordings. In this pilot study, we describe a first series of experiments to determine to what extent models of multilingual speech model can be used to detect languages that are phonetically/phonologically "similar".

Mots-clés : documentation computationnelle des langues, langues rares, apprentissage profond, typologie linguistique.

Keywords : computational language documentation, endangered languages, deep learning, linguistic typology.

1 Introduction

Les modèles de représentation de la parole multilingue appris de manière non supervisée par les réseaux de neurones comme `XLS-R` (Conneau *et al.*, 2021) et `HuBERT` (Hsu *et al.*, 2021) per-

Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain *(LIFT)* et Traitement Automatique des Langues *(TAL)*,
Marseille, France, 14 au 15 novembre 2022 , pages 67–73,

mettent de développer des systèmes de reconnaissance de la parole de bonne qualité à partir de très peu de données annotées. Ces technologies ouvrent de nombreuses possibilités pour la linguistique documentaire computationnelle : de nombreux travaux (Foley *et al.*, 2018 ; Anastasopoulos *et al.*, 2020 ; Partanen *et al.*, 2020), dont les nôtres (Macaire *et al.*, 2021 ; Guillaume *et al.*, 2022b,a), ont montré qu'il est possible de développer, pour des langues rares ou peu documentées, des systèmes de traitement de la parole pour faciliter le travail de transcription et d'annotation des linguistes de terrain. Nous souhaitons désormais pousser ces modèles encore plus loin en montrant que les représentations au cœur des systèmes de l'état de l'art peuvent apporter un autre type d'aide aux linguistes. Il s'agit de déceler automatiquement des propriétés importantes relatives à la typologie d'une langue (telles que : inventaire de phonèmes, indices de complexité phonologique et morphosyntaxique...), et, au moyen de ces propriétés, d'estimer le degré de proximité d'une langue avec telle ou telle autre.

Le travail exploratoire présenté ici est une première étape vers l'objectif ambitieux exposé ci-dessus. Nous souhaitons, dans un premier temps, déterminer dans quelle mesure les modèles de la parole multilingue peuvent être utilisés pour détecter des langues « similaires » au plan phonético-phonologique. Plus précisément, nous cherchons à savoir si les représentations de `XLS-R` permettent de distinguer deux langues et, plus généralement, dans quelle mesure deux langues « proches » (par exemple deux dialectes d'une même langue) auront des représentations plus proches que deux langues « éloignées ».

2 Estimer la similarité phonético-phonologique entre langues

Les modèles multilingues de représentation de la parole comme `XLS-R` ou `HuBERT` permettent de représenter un segment audio par un vecteur de taille fixe. Nous souhaitons vérifier que ces représentations encodent une information qui permet de caractériser la langue du segment.

Pour cela, nous nous inspirons des tests ABX utilisés en psychologie[1] ou dans l'analyse des représentations neuronales (de Seyssel *et al.*, 2022) ou pour l'apprentissage de systèmes d'identification du locuteur (Bredin, 2017) pour déterminer de manière complètement non supervisée (c'est-à-dire sans aucune annotation) si un modèle est capable de distinguer deux langues ou non. Ce test consiste à considérer les représentations construites par un modèle donné de deux segments audio de taille fixe A et X d'une langue donnée L_1 et un segment B d'une deuxième langue L_2, et à vérifier si la distance $d(A, X)$, mesurée par exemple par la distance cosinus, entre les deux vecteurs représentant A et X est plus petite que la distance $d(A, B)$ entre les représentations de deux segments dans des langues différentes.

Ce processus est répété pour un grand nombre de triplets (A, B, X) et nous calculons le score ABX correspondant à la proportion de triplets pour lesquels la condition $d(A, X) < d(A, B)$ est bien vérifiée. Plus ce score est grand, plus les représentations construites par le modèle pour les segments de la langue L_1 sont différents des segments de la langue L_2. Intuitivement, ce score ABX peut également être utilisé pour mesurer une distance entre langues : deux

1. Notons toutefois que, contrairement à l'approche adoptée ici, les tests ABX en psychologie sont souvent réalisés sur des « paires minimales », c'est-à-dire des exemples identiques à une caractéristique prêt (p. ex. un phonème).

langues proches seront plus confondues par le modèle et auront un score ABX plus petit que deux langues très différentes.

3 Expérience

Nous mettons en œuvre ce principe sur six langues de la collection `Pangloss` : quatre langues sino-tibétaines (trois langues de Chine : na de Yongning, na de Shekua, japhug ; et le thulung, parlé au Népal) ; le cèmuhî (langue austronésienne, Nouvelle Calédonie) ; et le nashta (langue indo-européenne, Grèce). Les deux premières langues sont des langues pouvant être qualifiées de « proches » (elles sont identifiées comme dialecte dans Pangloss).

Validation du protocole expérimental Dans une première expérience visant à valider le protocole expérimental décrit dans la section précédente, nous ne considérons que la paire de langue na du Yongning/japhug, les deux langues utilisées dans nos précédents travaux (Macaire *et al.*, 2021). Nous reportons, à la Table 1, l'évolution du score ABX en fonction de la longueur des segments considérés pour une paire de langue donnée. Notons que tous ces scores sont calculés de manière symétrique : il y a autant de triplets contenant deux enregistrements de la 1^{e} langue que de triplets contenant deux enregistrements de la 2^{e} langue. Intuitivement, plus les segments sont longs plus le modèle disposera d'information pour distinguer les langues et on peut donc s'attendre à ce que les scores ABX augmentent avec la durée des segments. Considérer des segments de longueurs différentes permet également de prendre en compte différents types d'information acoustiques : le modèle ne peut capturer que des informations « bas niveau » (p.ex. phonémiques) dans les segments les plus courts et des informations de plus « haut niveau » (p.ex. prosodiques) ne sont accessibles que dans des segments plus longs.

Les résultats de la Table 1 montrent clairement que les représentations construites automatiquement par `wav2vec` encodent des informations sur la langue : les scores ABX sont systématiquement au dessus de 50% et deviennent même très bon dès que les segments considérés sont suffisamment longs (plus de 5 s). Ce résultat est d'autant plus intéressant que les deux langues considérées ne sont pas présentent dans le corpus d'apprentissage : les représentations multilingues « découvertes » par `wav2vec` sont donc *génériques* et capables d'extraire des informations pertinentes y compris sur de nouvelles langues.

Mesure de la similarité entre langues Dans une seconde expérience, nous reportons à la Table 2 les scores ABX obtenus lorsque l'on cherche à distinguer des segments de 20 secondes pour différentes paires de langues qui peuvent être formées à partir des 6 langues considérées.

Ces résultats (même si encore préliminaires !) confirment que les représentations multilingues apprises automatiquement par `wav2vec` permettent de distinguer des langues, même si celles-ci ne sont pas présentes dans le corpus d'apprentissage. De manière plus intéressante, les performances semblent effectivement liées à une similarité entre langues puisque les scores ABX entre langues sino-tibétaines sont plus faibles qu'entre une langue sino-tibétaine et une autre langue (cèmuhî, nashta) : la moyenne de tous les scores ABX entre les langues

durée audio	score ABX
50 ms	59,2%
1 s	61,0%
5 s	82,2%
10 s	92,2%
20 s	94,3%
40 s	97,6%
50 s	97,0%

TABLE 1 – Évolution du score ABX pour la paire japhug, na de Yongning en fonction de la durée de l'échantillon.

	cèmuhî	nashta	na de Shekua	thulung	Japhug	na de Yongning
cèmuhî	—	74,6%	88,3%	62,7%	82,4%	80,6%
nashta	—	—	88,2%	62,3%	81,4%	88,4%
na de Shekua	—	—	—	79,8%	87,5%	87,2%
thulung	—	—	—	—	79,8%	76,2%
japhug	—	—	—	—	—	94,3%
na de Yongning	—	—	—	—	—	—

TABLE 2 – Score ABX pour différentes paires de langues. Les langues sino-tibétaine sont indiquées en orange, les autres langues en améthyste.

sino-tibétaine et les langues des autres familles est de $79,3\%$, alors qu'entre la moyenne de ces scores entre toutes les paires de langues sino-tibétaines est de $84,1\%$. Il est donc, de manière surprenante, plus difficile de distinguer les langues de deux familles différentes que les langues d'une même famille.

Pour corroborer les résultats de cette première expérience, nous nous sommes intéressés aux langues qianguiques, un sous-groupe de langues de la famille tibéto-birmane parlées en Chine, principalement sur le plateau de Qinghai. Les linguistes distinguent habituellement deux groupes de langues (Sims, 2016) : le qiang du Nord et le qiang du Sud. La collection Pangloss contient des enregistrements de 12 langues qiang, 5 identifiées comme des dialectes du qiang du Nord (luoduo, shuangliusuo, weicheng, waboliangzi et qugu) et 7 comme des dialectes du qiang du Sud (longxi, luobozhai, baishui, goukou, sanlong, heihu et baixi).

Nous avons considéré, pour chacune de ces 12 langues, 100 segments de 10 secondes, construit les représentations de ces segments à l'aide de `XSL-R` et mesuré la distance euclidienne entre toutes les paires de représentation. Nous avons alors calculé la moyenne des distances entre deux dialectes pour construire une matrice de distances entre dialectes et avons ensuite effectué un regroupement hiérarchique *(clustering)* de cette matrice à l'aide de l'algorithme du point le plus proche (Duda *et al.*, 2001) afin de construire un dendrogramme entre ces langues.

La hiérarchie et la matrice de similarité sont représentés à la Figure 1. Il apparaît clairement que les représentations construites par `XSL-R` permettent de retrouver la distinction entre les

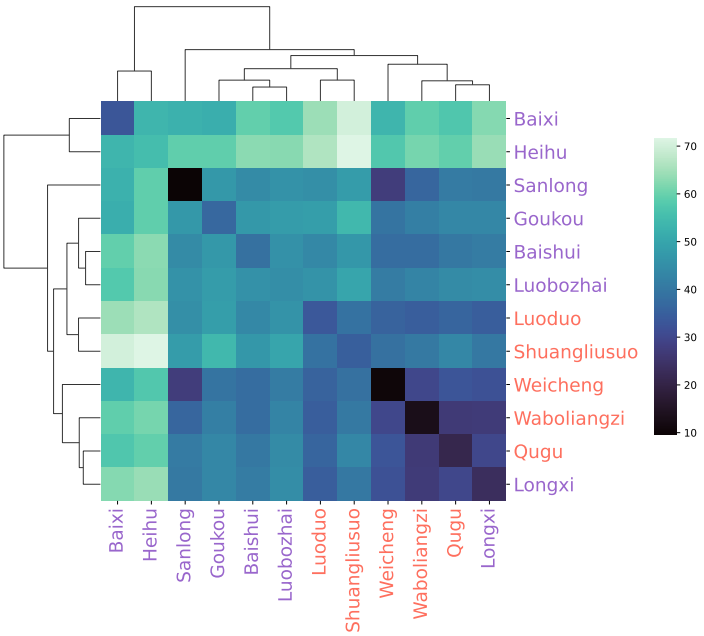


FIGURE 1 – Matrice de distances entre les langues qiang de la collection Pangloss et dendrogramme correspondant à la classification hiérarchique ascendante de celles-ci. Les dialectes du qiang du Nord sont représentés en orange et ceux du qiang du Sud en améthyste. Les dialectes ont été ordonnés automatiquement par similarité.

langues du qiang du Nord et celles appartenant au groupe du qiang du Sud (à une exception près). Ces représentations capturent donc bien des informations liées à la typologie des langues.

4 Conclusion

Les premiers résultats que nous présentons dans ce travail sont encourageants : s'ils soulèvent de nombreuses questions, ils ouvrent également la porte à de nombreuses perspectives particulièrement intéressantes. Ces résultats préliminaires devront toutefois être confirmés notamment pour garantir que la capacité du modèle à distinguer les langues repose bien sur des caractéristiques « linguistiques » et non uniquement sur des facteurs confondants (différences dans la manière dont les enregistrements ont été réalisés, locuteurs de genre différents, style de parole...).

Remerciements

Un grand merci à Alexis Michaud pour avoir initié ce travail et pour son soutien constant. Ses nombreux commentaires tant sur le fond que sur la forme ont grandement amélioré ce travail.

Nous remercions le CNRS/TGIR HUMA-NUM et le Centre de Calcul IN2P3 (Lyon - France) pour la fourniture des ressources informatiques et de traitement des données nécessaires à ce travail.

Ce travail a bénéficié du soutien financier de l'Agence Nationale de la Recherche (projet « La documentation computationnelle des langues à l'horizon 2025 » [ANR-19-CE38-0015-04] et Labex « Fondements empiriques de la linguistique » [ANR-10-LABX-0083]) ainsi que de l'Institut des Langues Rares (ILARA-EPHE).

Une partie importante des ressources linguistiques utilisées dans le présent travail a été collectée dans le cadre du projet « Corpus parallèles en langues himalayennes » [ANR-12-CORP-0006].

Références

Anastasopoulos A., Cox C., Neubig G. & Cruz H. (2020). Endangered languages meet modern NLP. In *Proceedings of the 28th International Conference on Computational Linguistics : Tutorial Abstracts*, p. 39–45, Barcelona, Spain (Online) : International Committee for Computational Linguistics. doi : 10.18653/v1/2020.coling-tutorials.7.

Bredin H. (2017). TristouNet : Triplet loss for speaker turn embedding. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, New Orleans, United States.

Conneau A., Baevski A., Collobert R., Mohamed A. & Auli M. (2021). Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, p. 2426–2430. doi : 10.21437/Interspeech.2021-329.

de Seyssel M., Lavechin M., Adi Y., Dupoux E. & Wisniewski G. (2022). Probing phoneme, language and speaker information in unsupervised speech representations. In *Interspeech 2022 - 23rd INTERSPEECH Conference*, Incheon, South Korea. hal : hal-03830470.

Duda R. O., Hart P. E. & Stork D. G. (2001). *Pattern Classification*. New York : Wiley, 2 édition.

Foley B., Arnold J., Coto-Solano R., Durantin G., Ellison T. M., van Esch D., Heath S., Kratochvíl F., Maxwell-Smith Z., Nash D., Olsson O., Richards M., San N., Stoakes H., Thieberger N. & Wiles J. (2018). Building speech recognition systems for language documentation : The CoEDL endangered language pipeline and inference system. In *Proceedings of the 6th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU 2018)*.

Guillaume S., Wisniewski G., Galliot B., Nguyễn M.-C., Fily M., Jacques G. & Michaud A. (2022a). Plugging a neural phoneme recognizer into a simple language model : a workflow for low-resource settings. In *Interspeech 2022 - 23rd Annual Conference of the International Speech Communication Association*, Proceedings of Interspeech 2022, Incheon, South Korea. doi : 10.5281/zenodo.5521111, hal : halshs-03625581.

Guillaume S., Wisniewski G., Macaire C., Jacques G., Michaud A., Galliot B., Coavoux M., Rossato S., Nguyễn M.-C. & Fily M. (2022b). Les modèles pré-entraînés à l'épreuve des langues rares : expériences de reconnaissance de mots sur la langue japhug (sino-tibétain). In *JEP 2022 - 34e Journées d'Études sur la Parole*, Actes des 34e Journées d'Études sur la Parole (JEP2022), Noirmoutier, France. hal : halshs-03625580.

Hsu W.-N., Bolte B., Tsai Y.-H. H., Lakhotia K., Salakhutdinov R. & Mohamed A. (2021). Hubert : Self-supervised speech representation learning by masked prediction

of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, **29**, 3451–3460. DOI : 10.1109/TASLP.2021.3122291.

Macaire C., Wisniewski G., Guillaume S., Galliot B., Jacques G., Michaud A., Rossato S., Nguyễn M.-C. & Fily M. (2021). Spécialisation de modèles neuronaux pour la transcription phonémique : premiers pas vers la reconnaissance de mots pour les langues rares. In *Journées scientifiques du Groupement de recherche "Linguistique informatique, formelle et de terrain" (GDR LIFT)*, Journées GDR LIFT 2021, Grenoble, France. HAL : halshs-03475443.

Partanen N., Hämäläinen M. & Klooster T. (2020). Speech recognition for endangered and extinct samoyedic languages. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, p. 523–533, Hanoi, Vietnam : Association for Computational Linguistics.

Sims N. (2016). Towards a more comprehensive understanding of Qiang dialectology. *Language and Linguistics*, **17**(3), 351–381.

Documentary Research in Natural Language (D.R.N.L.) : Plateforme d'accès numérique aux archives documentaires en langage naturel

Ying ZHANG[1] Matthieu PETIT GUILLAUME[1] Aurélien KRAUTH[1]
(1) Leviatan, 725 Boulevard Robert Barrier, 73100 Aix-les-Bains, France
y.zhang@leviatan.fr, matthieu@leviatan.fr, aurelien@leviatan.fr

RÉSUMÉ

Nos travaux de recherche sont motivés par un besoin industriel consistant initialement à la gestion, au stockage et à l'accès à d'anciens journaux et magazines archivés. Notre partenaire industriel possède plusieurs téraoctets de magazines et de journaux. Ces documents sont rédigés dans différentes langues (français, russe, portugais, etc.), répartis dans plusieurs dossiers représentant chacun un type de magazine précis. Ils sont stockés en format PDF et JPG. Dans le cadre de cet article, nous avons centré notre recherche sur le traitement des documents français. Nous proposons une plateforme D.R.N.L. (Documentary Research in Natural Language) permettant le traitement, le stockage et l'accès à des archives documentaires avec quatre composants principaux : 1. Prétraitement des magazines, 2. Stockage des données, 3. Filtrages des documents à analyser pour une question posée et 4. Inférence de requête.

ABSTRACT

Documentary Research in Natural Language (D.R.N.L.): Platform for digital access to documentary archives in natural language

Our research was motivated by an industrial need. It initially involves managing, storing and accessing old archived newspapers and magazines. Our partners own terabytes of magazines and newspapers. These documents are written in different languages (French, Russian, Portuguese, etc.) divided into several folders, each representing a specific type of magazine. They are stored in PDF and JPG format. In the context of this article, we have focused our research on the processing of French documents. We offer a D.R.N.L. (Documentary Research in Natural Language) allowing the processing, storage and access of documentary archives in four main components: 1. pre-processing of magazines, 2. data storage, 3. filtering of documents to be analyzed for a question asked and 4 query inferences.

MOTS-CLÉS : Compréhension automatique de texte, Système de questions-réponses, Analyse automatique des archives documentaires, Moteur de recherche de données dédiées

KEYWORDS: Machine reading comprehension, Question answering system, Automatic analysis of documentary archives, Dedicated data search engine

Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain *(LIFT) et* Traitement Automatique des Langues *(TAL),*
Marseille, France, 14 au 15 novembre 2022 , pages 74–83, ©2022 CNRS

1 Introduction du projet D.R.N.L.

A l'heure d'internet, il est de plus en plus facile et accessible de rechercher de l'information sur de nombreux de sujets.

Les archives documentaires et notamment celles générées par la presse spécialisée, jouent un rôle important chez les professionnels, même si celles-ci ont opéré leur transformation vers le numérique. Ainsi entre 1990 et 2004, la diffusion annuelle au format papier a reculée de près de 38% (Tessier, 2007).

Mais que deviennent ces archives documentaires et notamment les anciens numéros de magazines et de journaux spécialisés ? Ceux-ci regorgent d'informations riches et précieuses dont la numérisation représente une solution efficace de stockage et un moyen rapide de recherche d'informations précises et pertinentes mis en œuvre au travers d'une interface homme machine (IHM).

Notre partenaire a stocké plusieurs téraoctets de magazines et de journaux. Ces documents sont rédigés dans différentes langues (français, russe, portugais, etc.). Dans la première phase de ce projet, nous avons centré notre recherche sur le traitement des documents français. Cela implique 1.1 téraoctets de magazines et journaux français.

Dans cet article, nous proposons une nouvelle plateforme D.R.N.L. (Documentary Research in Natural Language) permettant le traitement, le stockage et l'accès à des archives documentaires. Notre plateforme a été séparée en deux parties : 1. Développement de l'IHM et 2. Recherche en TAL (Traitement Automatique des Langues). Dans cet article, nous ne présentons pas la réalisation de l'IHM ni la gestion des utilisateurs.

L'objectif de ces travaux de recherche est de stocker des données numériques standardisées et normalisées, de rechercher des données à l'aide de mots-clés, d'inférer des requêtes et de proposer des réponses.

La plateforme est implémentée sous quatre composants principaux :

1. Prétraitement des magazines,
2. Stockage des données,
3. Filtrage des documents à analyser pour une question posée, et
4. Inférence de requête.

Cet article est organisé de la façon suivante. Nous présentons les difficultés rencontrées et les solutions retenues pour les composants principaux ci-dessus dans les sections 2 à 5. Ensuite nous présentons les expérimentations et les déploiements. Enfin, nous concluons et donnons quelques perspectives.

2 Prétraitement des magazines

Les archives documentaires sont stockées en PDF ou en image. Ce composant permet d'unifier les archives documentaires et de nous envoyer les sorties en deux catégories.

La première catégorie contient les sorties des informations des paragraphes internes, tels que le texte brut, les parties du discours, les lemmes, la langue, la position du paragraphe dans la page, l'extraction des mots-clés et la transformation du plongement lexical (word embeddings) etc.

La deuxième catégorie contient les sorties des informations des paragraphes externes, par exemple, le contexte du paragraphe (les informations du paragraphe précédent et du paragraphe suivant), le nom du document, le numéro de la page du paragraphe, le nombre de pages total, et l'année de parution etc.

2.1 Problématiques

Les documents sont hétérogènes. Les mises en page sont incohérentes, même pour différentes pages du même document. Dans un même document, nous observons des pages à deux colonnes, des pages à trois colonnes, et des pages divisées en partie haute, centrale et basse etc. Le nombre et la largeur des colonnes ne sont pas identiques avec de nombreux petits blocs d'annonces. Les en-têtes et les pieds des pages amènent également beaucoup d'informations redondantes. Il existe des magazines bilingues (en français-anglais ou en français-allemand). Ces éléments augmentent la difficulté de la normalisation.

2.2 Solutions

Notre solution est principalement divisée en quatre étapes. La première étape consiste en une analyse d'OCR (Smith, 2007), la détection de position de chaque texte et une génération des PDF éditables. Ensuite nous avons réalisé une manipulation et analyse d'objets géométriques basés sur des coordonnées cartésiennes. Nous fusionnons deux textes dans un paragraphe, si la distance entre ces deux textes consécutifs est inférieure à 5 unités de longueur (Gillies, 2013). La sortie de cette étape est un ensemble de paragraphes.

La deuxième étape est le nettoyage. Nous nettoyons toutes les en-têtes et les pieds de page en se basant sur la position et la longueur du texte. Le projet actuel ne permet pas de nettoyer le contenu de l'annonce. Nous discuterons de ce point plus loin dans la section perspectives.

Dans la troisième étape, nous utilisons SpaCy (DataCamp, 2020) afin d'effectuer une analyse linguistique sur chaque paragraphe, comprenant principalement la détection de la langue, la

tokenisation et la lemmatisation. Enfin nous utilisons Yake (Campos et al., 2020) afin de faire une extraction de mots-clés.

Dans la dernière étape, nous transformons les textes en plongement lexical (word embeddings) en utilisant un modèle de langage pré-entraîné. Bien que dans la première phase du projet, nous ne traitions que des documents français, pour une mise à l'échelle plus facile vers des plateformes multilingues, nous choisissons le modèle multilingue de l'équipe Google (Chidambaram et al., 2019; Yang et al., 2020) pour mettre en œuvre cette étape.

3 Stockage des données

Le stockage des données est séparé en deux parties. 1. Les documents originaux et les PDF éditables sont stockés dans un bucket AWS S3 (AWS, 2006), 2. Les données numériques sont stockées dans un index Elasticsearch (Elasticsearch, 2018). Les PDF éditables sont utilisés pour l'affichage sur l'IHM. Les données numériques sont utilisées pour le calcul de la réponse.

À ce stade de cet article, nous avons stocké 1.1 téraoctets d'archives documentaires en français dans le S3 et plus de 500 000 paragraphes dans l'index d'Elasticsearch.

4 Filtrages des paragraphes à analyser pour une question posée

Le système D.R.N.L. permet de traiter deux types de requête : 1. Recherche à l'aide de mot(s)-clé(s) et 2. Inférence d'une question posée.

Pour traiter le premier type de requête, nous utilisons le résultat de l'extraction de mots-clés stockés dans les données numériques. Si ce résultat est vide pour un mot-clé indiqué par l'utilisateur, nous utilisons le résultat de lemmatisation stocké dans les données numériques (voir la section 2.2). Nous ne détaillons pas cette implémentation. Dans cette section et la suivante, nous nous concentrerons sur le cœur de notre recherche, à savoir comment recommander des réponses à une question donnée dans le contexte de données massives et de contraintes par les besoins industriels.

4.1 Problématiques

Dans la phase initiale du projet, nous avons réalisé un prototype avec moins de 500 paragraphes stockés dans l'index d'Elasticsearch. Dans ce prototype, il n'y a pas d'étape de filtrage, et nous obtenons un bon fonctionnement.

Une fois que la quantité de données dans Elasticsearch augmente, des problèmes de rapidité et de passage à l'échelle sont apparus. Étant donné que notre processus d'inférence utilise un modèle MRC

(Machine Reading Comprehension) affiné et basé sur CamemBERT (Martin et al., 2020), son fonctionnement consomme beaucoup de ressources de calcul. Nous le présenterons dans la section 5. Avec notre test et calcul, une analyse de 500 000 paragraphes sans filtrage, prend environ 45 minutes afin de recevoir les résultats. Ce test est basé sur un déploiement sur machine GPU NVIDIA Tesla V100.

D'autre part, la précision des réponses a également chuté de manière significative. Nous voulons souligner notre observation sur le traitement des noms propres. Par exemple, pour la question « *Qui est le président de Dior ? »*, en supposant que nous ayons les deux textes suivants : 1. « *Le président de Chanel est Bruno Pavlovsky.* » et 2. « *La nomination inattendue de Pietro Beccari à la tête de Dior récemment.* », le modèle MRC préfère recommander « *Bruno Pavlovsky* » comme réponse à la question. La raison est que la parenté sémantique de ces deux mots (*Dior* et *Chanel*) est trop élevée, autrement dit, les word embeddings de ces deux mots sont très similaires. Au contraire, la relation sémantique entre « *La nomination inattendue de Pietro Beccari à la tête de Dior récemment.* » et « *Qui est le président de Dior ?* » est plus éloignée que la relation sémantique entre « *Le président de Chanel est Bruno Pavlovsky.* » et « *Qui est le président de Dior ?* ». Lorsque le nom propre dans la question fait référence à une marque ou à un objet moins connu, comme un modèle précisé d'une marque de téléphone mobile, ou un nouveau parfum, le retour du système sera encore moins pertinent. Nous observons que les word embeddings de ces noms propres sont convertis en un vecteur fixe par le modèle.

En raison des contraintes industrielles, nous ne pouvons pas construire une base de connaissance afin de traiter ce problème. Face à une grande quantité de données, il est extrêmement difficile pour notre prototype de faire des inférences correctes sur des questions avec un nom propre.

Pour les raisons ci-dessus, nous avons ajouté le filtrage comme étape essentielle. L'objectif de cette étape : étant donnée une question posée par l'utilisateur, nous avons utilisé plusieurs stratégies de filtrages afin de récupérer les premiers 1000 paragraphes les plus pertinents.

4.2 Solutions

Lors d'une recherche Elasticsearch, un score est calculé pour chaque document dans le résultat. Ce score représente la pertinence du document afin de pouvoir en classer les résultats.

Une question est généralement relativement courte, mais un paragraphe est régulièrement long. En raison d'une grande différence de longueur entre la question et le paragraphe, Elasticsearch donne des scores plus élevés aux paragraphes courts. Par exemple, pour une question « *Quand la société Dior a-t-elle été créée ?* », nous avons deux textes, le premier texte est un titre d'un article « *Société Dior et sa création* », le deuxième texte est un paragraphe décrivant l'historique de croissance de la société Dior. Le score du premier texte est plus élevé que le deuxième texte. C'est parce que Elasticsearch calcul son score selon la fréquence du terme, la fréquence inverse du terme et la longueur du champ (Denton, 2017).

Cette étape de filtrage prend donc en compte trois axes dans le calcul :

1. La proposition du moteur de recherche d'Elasticsearch et la longueur du texte. Si nous avons assez de paragraphes candidats, nous ignorons les paragraphes courts (le nombre de tokens est inférieur à 20).
2. La cohérence du nom propre. S'il existe un nom propre dans la question, celui-ci doit également exister dans le paragraphe courant.
3. Le filtrage de la similarité cosinus basé sur le word embeddings entre les différents paragraphes candidats. Si deux paragraphes candidats ont une similarité très élevée, nous n'en prenons qu'un pour l'analyse. C'est principalement parce qu'un même contenu ou un même sujet est rapporté à plusieurs reprises par différents éditeurs ou journaux.

Enfin, nous prenons les premiers 1000 paragraphes les plus pertinents comme les paragraphes candidats, analysés par le modèle MRC.

5 Inférence de requête

Du point de vue du domaine de la connaissance, un système de questions-réponses peut être divisé en deux types : « domaine fermé » et « domaine ouvert ». Si les archives documentaires appartiennent toutes au même domaine, alors l'approche d'une ontologie dédiée et une base de connaissance pourront grandement améliorer la précision des réponses (Lopez et al., 2007; Otegi et al., 2015; Siciliani, 2018; Franco et al., 2020).

Dans le cadre du projet D.R.N.L., une grande quantité de données brutes hétérogènes couvre de nombreux domaines tels que la politique, les affaires, le divertissement, et la mode etc. Avec de nombreuses contraintes industrielles, nous devons développer une solution rapide et efficace. Enfin, nous nous concentrons sur l'étude du modèle MRC (Machine Reading Comprehension).

5.1 État de l'art

Le MRC est un sujet très important dans le domaine TALN (Traitement Automatique du Langage Naturel). Des recherches récentes sur l'utilisation des modèles de langue contextualisés pré-entraînés avec des architectures à base de Transformer (Vaswani et al., 2017) ont obtenu un grand succès dans de nombreuses tâches de TALN.

CamemBERT (Martin et al., 2020) est considéré comme l'un des meilleurs modèles français. Nous utilisons ce modèle comme le modèle de langue contextualisé pré-entraîné pour commencer notre expérimentation.

5.2 Modèle MRC ajusté

L'équipe Google AI Langue a expliqué la méthode d'ajustement d'un système MRC basé sur le modèle de langue BERT (Devlin et al., 2019) en utilisant les jeux de données SQuAD v1.1 (Rajpurkar et al., 2016) et SQuAD v2.0 (Rajpurkar et al., 2018). Nous n'aborderons pas le détail de cette méthode dans cet article.

Nous avons bien respecté les consignes précisées dans (Devlin et al., 2019) et avons gardé les valeurs par défaut des hyper-paramètres proposées par Transformer (Huggingface, 2020) pour ajuster notre modèle MRC en utilisant trois jeux de données publiques et un jeu de données privées de notre partenaire. Les trois jeux de données en source ouverte proviennent du projet Piaf (Bras, 2019), du projet FQuAD (D'Hoffschmidt et al., 2020) et du projet French-SQuAD (Kabbadj, 2018). Le F1-score atteint une valeur moyenne de 80.6 sur cet ensemble de jeux de données.

6 Déploiements et expérimentations

Le système D.R.N.L. est un système qui permet d'accéder au service en temps réel, par conséquent, la performance de ce système est très importante. En termes de déploiement et d'optimisation de code, nous avons fait plusieurs tentatives. Dans ce système, la partie la plus gourmande en ressources est l'analyse du modèle MRC des 1000 paragraphes les plus pertinents (voir la section 4.2) pour une question donnée.

Nous transformons les analyses en forme d'itération *question1 : [text1, text2, text3…text1000]* vers les analyses en forme *[question1 : text1, question1 : text2, question1 : text3…question1 : text1000]*, en utilisant le traitement « batching » du pipeline de Transformer (Huggingface, 2021). Pour garantir un temps de traitement raisonnable, cette partie est déployée sur une machine GPU NVIDIA Tesla V100.

Nous avons mis en place d'autres solutions d'optimisation du programme. Par exemple, nous pré-analysons et stockons les word embeddings et les résultats d'analyse linguistique dans l'index d'Elasticsearch. Ces optimisations ont permis une importante amélioration du temps de traitement, qui est passé de dizaines de secondes, voire 1 minute, à environ 4-5s.

Nous avons testé ce système avec 4050 questions pré-annotées. 3702 questions présentent des réponses et 348 questions liées aux bons documents mais n'ayant pas de réponses précises. Nous proposons au maximum 10 réponses pour chaque question selon l'ordre du score de la fiabilité.

Dans les 3702 questions avec réponses, nous avons 3077 bonnes réponses trouvées. Parmi ces 3077 bonnes réponses, 2906 réponses ont reçu un score de fiabilité élevé du MRC (>0.2), 171 réponses ont reçu un score de fiabilité faible (<0.2). Le système a proposé 355 mauvaises réponses mais trouvées

dans le bon document. Enfin, le système a proposé 271 mauvaises réponses qui ne sont pas présentes dans les bons documents.

Dans les 348 questions sans réponses, nous avons retrouvé 160 bons documents (79 questions ont reçu une réponse avec un score de fiabilité élevé et 81 questions ont reçu une réponse avec un score de fiabilité faible). 188 questions n'ont pas pu être rattachées au bon document.

7 Conclusions et perspectives

Dans cet article, nous présentons une nouvelle plate-forme D.R.N.L. basée sur les technologies les plus récentes en TALN, permettant de stocker et d'accéder de façon plus efficace et plus durable aux archives documentaires.

Les perspectives de cette recherche sont multiples et concernent aussi bien le court terme que le long terme. En ce qui concerne le court terme, il s'agit d'un passage à échelle en multilingue. Pour le long terme, nous pouvons distinguer deux perspectives.

Notre IHM D.R.N.L. permet non seulement de présenter les dix premiers résultats d'une question, mais également de gérer un espace membre dédié afin que les utilisateurs puissent partager et contribuer aux contenus les plus pertinents associés aux sujets donnés. Nous avons prévu de construire un système de recommandation en utilisant les contributions des utilisateurs.

La deuxième perspective a été mentionnée aux sections 2.1 et 2.2. Il s'agit d'un modèle de classification de textes afin d'identifier les annonces. Aujourd'hui les annonces nous amènent beaucoup de bruits dans le stockage. Afin d'entraîner ce modèle, nous voulons utiliser la méthode présentée dans (Sun et al., 2019). Nous commencerons à annoter les données dans un proche avenir.

FIGURE 1 : Interface d'inférence pour la question « Qui est le directeur marketing de Darty ? »

Références

AWS. (2006). Amazon Simple Storage Service. Available at: https://aws.amazon.com/fr/s3/

BRAS, M. (2019). Piaf. Available at: https://www.etalab.gouv.fr/ia-decouvrez-et-participez-au-projet-piaf-pour-des-ia-francophones

CAMPOS, R., MANGARAVITE, V., PASQUALI, A., JORGE, A., NUNES, C., & JATOWT, A. (2020). YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, *509*, 257–289. DOI : https://doi.org/10.1016/j.ins.2019.09.013

CHIDAMBARAM, M., YANG, Y., CER, D., YUAN, S., SUNG, Y. H., STROPE, B., & KURZWEIL, R. (2019). Learning cross-lingual sentence representations via a multi-task dual-encoder model. *ACL 2019 - 4th Workshop on Representation Learning for NLP, RepL4NLP 2019 - Proceedings of the Workshop*, 250–259. DOI : https://doi.org/10.18653/v1/w19-4330

D'HOFFSCHMIDT, M., VIDAL, M., BELBLIDIA, W., & BRENDLÉ, T. (2020). FQuAD: French question answering dataset. DOI : https://doi.org/10.18653/v1/2020.findings-emnlp.107

DATACAMP. (2020). spaCy. *Python Cheat Sheet*. Available at: https://www.datacamp.com/cheat-sheet/spacy-cheat-sheet-advanced-nlp-in-python

DENTON, A. (2017). *Making your search not suck with Elasticsearch — Part 6: Totally irrelevant*.

DEVLIN, J., CHANG, M. W., LEE, K., & TOUTANOVA, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*. DOI : https://doi.org/10.18653/v1/N19-1423

ELASTICSEARCH. (2018). Elasticsearch. *International Journal of Modern Trends in Engineering & Research*, *5*(5), 23–28.

FRANCO, W., VIKTOR, C., OLIVEIRA, A., MAIA, G., BRAYNER, A., VIDAL, V. M. P., CARVALHO, F., & PEQUENO, V. M. (2020). Ontology-based question answering systems over knowledge bases: A survey. *ICEIS 2020 - Proceedings of the 22nd International Conference on Enterprise Information Systems*. DOI : https://doi.org/10.5220/0009392205320539

GILLIES, S. (2013). *The shapely user manual, Version 1.3*. December 31, 2013.

HUGGINGFACE. (2020). Fine-tuning BERT on SQuAD1.0. https://huggingface.co/transformers/v2.8.0/examples.html#squad

HUGGINGFACE. (2021). Pipeline batching. https://huggingface.co/docs/transformers/main_classes/pipelines#pipeline-batching

KABBADJ, A. (2018). Something new in French Text Mining and Information Extraction (Universal Chatbot): Largest Q&A French training dataset (110 000+).

LOPEZ, V., UREN, V., MOTTA, E., & PASIN, M. (2007). AquaLog: An ontology-driven question answering system for organizational semantic intranets. *Web Semantics*. DOI : https://doi.org/10.1016/j.websem.2007.03.003

MARTIN, L., MULLER, B., SUAREZ, P. J. O., DUPONT, Y., ROMARY, L., DE LA CLERGERIE, É. V., SEDDAH, D., & SAGOT, B. (2020). CamemBERT: A tasty French language model. DOI : https://doi.org/10.18653/v1/2020.acl-main.645

OTEGI, A., ARREGI, X., ANSA, O., & AGIRRE, E. (2015). Using knowledge-based relatedness for information retrieval. *Knowledge and Information Systems*. DOI : https://doi.org/10.1007/s10115-014-0785-4

RAJPURKAR, P., JIA, R., & LIANG, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*. DOI : https://doi.org/10.18653/v1/p18-2124
RAJPURKAR, P., ZHANG, J., LOPYREV, K., & LIANG, P. (2016). SQuad: 100,000+ questions for machine comprehension of text. *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*. DOI : https://doi.org/10.18653/v1/d16-1264
SICILIANI, L. (2018). Question answering over knowledge bases. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. DOI : https://doi.org/10.1007/978-3-319-98192-5_47
SMITH, R. (2007). Tesseract OCR Engine. *Lecture. Google Code. Google Inc*.
SUN, C., QIU, X., XU, Y., & HUANG, X. (2019). How to Fine-Tune BERT for Text Classification? *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. DOI : https://doi.org/10.1007/978-3-030-32381-3_16
TESSIER, M. (2007). La presse au défi du numérique. In *RAPPORT AU MINISTRE DE LA CULTURE ET DE LA COMMUNICATION*.
VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, Ł., & POLOSUKHIN, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*. DOI : https://doi.org/10.48550/arXiv.1706.03762
YANG, Y., CER, D., AHMAD, A., GUO, M., LAW, J., CONSTANT, N., ABREGO, G. H., YUAN, S., TAR, C., SUNG, Y., STROPE, B., & KURZWEIL, R. (2020). *Multilingual Universal Sentence Encoder for Semantic Retrieval*. 87–94. DOI : https://doi.org/10.18653/v1/2020.acl-demos.12

Évaluation de techniques non supervisées pour l'assistance à l'annotation manuelle de textes

Kévin Deturck[1] Hugo Lafayette[2] Bénédicte Diot-Parvaz Ahmad[1] Ilaine Wang[1]
Afala Phaxay[1] Damien Nouvel[1]
(1) Inalco – Ertim, 2, rue de Lille, 75007 Paris, France
(2) Kairntech, 29, chemin du vieux chêne, 38240 Meylan, France
kevin.deturck@inalco.fr, hugo.lafayette@kairntech.com,
benedicte.parvazahmad@inalco.fr, ilaine.wang@inalco.fr,
afala.phaxay@inalco.fr, damien.nouvel@inalco.fr

RÉSUMÉ

La production de données annotées est critique pour l'élaboration de nombreux systèmes de TAL, avec des problèmes liés à l'expertise des annotateurs et à l'organisation de leur travail dans l'optique d'obtenir des annotations de qualité. Nous évaluons l'impact de techniques d'assistance non supervisées pour la catégorisation manuelle de textes, lors de trois campagnes d'annotation d'articles journalistiques impliquant une vingtaine d'annotateurs à travers trois langues, l'hindi, le mandarin et l'arabe. Cette évaluation est quantitative, considérant le niveau de qualité ainsi que la vitesse de l'annotation, et qualitative en intégrant les retours d'expérience des annotateurs. Elle a permis la mise en œuvre innovante d'assistances basées sur des combinaisons d'approches standards et nouvelles ainsi que d'éprouver une méthode pour l'encadrement de campagnes d'annotation multi-annotateurs.

ABSTRACT

Evaluation of unsupervised techniques for assisted manual text categorisation

The production of annotated data is critical for the development of many NLP systems, with problems related to the expertise of the annotators and the organisation of their work in order to obtain quality annotations. We evaluate the impact of unsupervised techniques for assisted manual categorisation of journalistic articles by means of three annotation campaigns involving twenty annotators, each one being dedicated to one language among Hindi, Mandarin and Arabic. The evaluation is both quantitative, considering the level of quality as well as the speed of the annotation, and qualitative by integrating the annotators' feedback. It allowed for an innovative implementation of assistance tools based on standard and new approaches as well as the testing of a method for managing multi-annotator annotation campaigns.

MOTS-CLÉS : campagne d'annotation, classification de texte, TAL, zeroshot, bertopic

KEYWORDS: annotation campaign, NLP, text classification, zeroshot, bertopic

Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain *(LIFT) et* Traitement Automatique des Langues *(TAL),*
Marseille, France, 14 au 15 novembre 2022 , pages 84–92,

1 Introduction

La collecte et l'exploitation des données langagières sont devenues aujourd'hui des enjeux stratégiques dans la vie économique et sociale. Les apports de l'intelligence artificielle au TAL ont transféré les difficultés de mise au point d'un projet sur la production de données annotées de qualité, préalable indispensable à la création de modèles capables d'effectuer une tâche de TAL efficacement.

Le travail que nous présentons ici s'inscrit dans le cadre du projet VITAL (Valorisation de l'Innovation pour le Traitement Automatique des Langues), qui consiste à étudier de nouveaux outils pour produire des données annotées avec le meilleur compromis entre qualité, coût et temps de développement. Ce projet porte distinctement sur trois phases d'une campagne d'annotation : l'amorçage, la consolidation et la clôture.

C'est un partenariat industrie-recherche. Kairntech développe et commercialise depuis trois ans une plateforme de TAL qui permet, par une interface graphique simple d'utilisation, de créer des jeux de données annotées puis de les utiliser en implémentant, évaluant et en maintenant des modèles d'IA. Pour Vital, Kairntech développe des outils d'assistance à l'annotation manuelle sous forme de prototypes dans leur plateforme d'annotation. Le laboratoire de recherche ERTIM de l'Inalco, spécialiste en TAL multilingue, gère les campagnes d'annotation pour Vital. Le projet est financé par le dispositif RAPID, porté par l'agence de l'innovation de défense.

Ce travail concerne l'évaluation d'outils d'assistance à l'amorçage d'une tâche de classification manuelle de textes en hindi, mandarin et arabe, avec des techniques non supervisées. Dans la suite, nous positionnons ce travail par rapport à l'état de l'art en section 2, puis, nous décrivons les campagnes d'annotation organisées et les techniques d'assistance étudiées respectivement en sections 3 et 4, et enfin, nous présentons l'analyse et les résultats en section 5 avant de conclure en section 6.

2 État de l'art

2.1 Campagne d'annotation : organisation et analyse

Il a été démontré qu'une phase de formation des annotateurs était un levier fondamental pour maximiser la qualité de leur travail (Bayerl & Paul, 2011), avec en support, un guide d'annotation pour la cohérence des annotations (Alex et al., 2006). Ces composantes s'insèrent dans un processus d'annotation dont il existe deux grands types (Fort, 2012) : un dit « traditionnel », avec une seule itération des phases de préparation, d'annotation et d'analyse, et un autre, plus récent, dit « agile », avec plusieurs itérations progressives de ces phases. Ces révisions font suite aux retours d'expérience des annotateurs ou à l'analyse de leurs désaccords d'annotation. Ces désaccords peuvent être résolus par un expert de la tâche ou par les annotateurs eux-mêmes. Nos campagnes intègrent ces composantes pour donner un cadre standard à notre étude.

Pour évaluer la qualité des annotations produites, une approche classique est de mesurer l'accord inter-annotateur (Artstein & Poesio, 2008). Le critère de qualité sous-jacent est la reproductibilité du schéma d'annotation par plusieurs annotateurs (Krippendorff, 2004) ; un schéma d'annotation est d'autant plus fiable qu'il est reproduit par plusieurs annotateurs. Ce type d'approche a notamment l'avantage de ne pas nécessiter de référence, souvent indisponible, mais ne permet pas véritablement de dire si un ensemble d'annotations est correct, à l'inverse des métriques de comparaison à une référence, comme la F-mesure ou le « Slot error rate », moins représentées. Nous utiliserons la F-mesure parce que le protocole d'évaluation choisi nous a permis d'acquérir un corpus de référence.

2.2 Assistance à l'amorçage de la catégorisation manuelle de texte

« Zero-shot » est la dénomination d'un ensemble de techniques d'apprentissage se proposant de résoudre une tâche sans avoir d'exemple à disposition, puisqu'une description des différentes classes attendues suffit (Winata et al. 2021). Nous en avons choisi une implémentation à base de transformeurs et de modèles d'infèrence en langage naturel, adaptée à notre cas d'usage de classification de documents. Les performances de modèles multilingues comme mBERT sont réputées bonnes sur des langues dotées, comme l'arabe et le mandarin (Wu & Dredze. 2020). BERTopic, qui exploite ces modèles multilingues, semble apporter des améliorations par rapport à l'état de l'art de la classification thématique de documents (Egger & Yu 2022). Il nous semble alors opportun de tester l'usage et les performances de BERTopic dans le cadre de VITAL.

Les outils de recherche par mots-clés ainsi que par facettes sont assez standards pour l'exploration de corpus, notamment dans le domaine des humanités numériques (Schnober & Gurevych, 2015). Cependant, l'étude de leur impact sur une tâche de classification manuelle n'est pas documentée à notre connaissance. Par ailleurs, certaines plateformes d'étiquetage manuel, comme Prodigy et TagTog, ne fournissent pas ou ne mettent pas en avant ces outils d'assistance. Nous souhaitons évaluer ce qu'ils apportent en pratique durant des campagnes d'annotation.

3 Les campagnes d'annotation

3.1 La tâche d'annotation

La tâche d'annotation du lot 1 consiste à catégoriser en thème chacun des articles de journaux, tous écrits dans une même langue, en choisissant une seule étiquette de catégorie parmi un ensemble d'étiquettes imposé. Cela correspond à de la classification « multiclasse » (Rifkin, 2008) parce qu'il y a plusieurs étiquettes de catégorie possibles et une seule doit être choisie. Outre les étiquettes de catégorie, il y a aussi la possibilité de sélectionner une étiquette intitulée « Incertitude » quand aucune

étiquette de catégorie ne convient ou lorsqu'il y a un doute sur le choix d'une étiquette de catégorie, ceci afin de tracer les difficultés rencontrées.

La tâche d'annotation requiert de repérer le sujet majeur d'un document pour lui associer une catégorie. Ce travail est difficile parce que les articles journalistiques mêlent parfois plusieurs domaines de connaissance, ce qui laisse place à l'interprétation et donc à la subjectivité pour déterminer ce qui est majeur, compliquant l'obtention d'annotations cohérentes. La tâche est par ailleurs complexifiée par la nécessité de considérer l'ensemble du texte d'un article (par une attention soutenue) avant de prendre une décision pour l'annoter, ainsi que par le besoin de disposer d'éléments de connaissance sur les domaines abordés pour éclairer la décision.

3.2 Les étiquettes de catégorie

Hindi	Mandarin	Arabe
7 étiquettes	7 étiquettes	6 étiquettes
- Sport - Économie - Divertissements et médias - Société - Monde - Politique - Sciences et techniques	- Sport - Divertissements - Société - Monde - Culture - Santé et bien-être - Anti-corruption	- Sport - Économie - Politique internationale - Politique d'Oman - Art et culture - Religion

TABLE 1 : Les étiquettes de catégorie par campagne

Les journaux sources des corpus avaient leur catégorisation respective des documents. Cependant, nous avons choisi de ne pas reprendre les catégorisations originelles car nous avions observé qu'elles manquaient parfois de cohérence : par exemple, dans le corpus en hindi, un article sur le discours du Premier Ministre indien en Chine était catégorisé tantôt en « International », tantôt en « Politique ».

Pour définir les jeux d'étiquettes (cf. Table 1), nous avons fait appel à un « référent langue » par campagne : c'est une locutrice ou un locuteur distinct des annotateurs. Le travail du référent langue a été de faire une pré-annotation d'une petite partie du corpus (d'une cinquantaine à une centaine de documents), en utilisant au départ le jeu de catégories originel et en ayant la liberté de le modifier. Ce travail de pré-annotation par les référents langue fut à la base de la rédaction du guide d'annotation de chacune des campagnes.

3.3 Les corpus

Nous avons constitué un corpus d'articles journalistiques pour chacune des trois campagnes. Le corpus pour la campagne sur l'arabe provient d'un corpus pré-existant, « El-Watan 2004 » (Abbas et al., 2005), c'est le seul à être issu d'une source unique, le journal arabe « El Watan », en édition omanaise et en ligne. Nous avons extrait les deux autres corpus à partir de plusieurs journaux en ligne, sept pour le corpus en hindi et trois pour le corpus en mandarin.

Pour favoriser la comparabilité des campagnes, nous avons choisi de constituer des corpus de tailles similaires Le nombre de 900 documents a été initialement choisi afin de maximiser le nombre de documents annotés pendant chaque session d'annotation. Le corpus en arabe a un nombre un peu plus élevé de documents (954) parce que nous avions observé qu'il était judicieux de proposer un peu plus de documents par session d'annotation pour laisser plus de place à d'éventuels contrastes concernant la vitesse d'annotation entre les annotateurs, en particulier selon l'utilisation ou non d'une assistance.

3.4 Les annotatrices et annotateurs

Campagne	Hindi	Mandarin	Arabe
Nombre	6	5	10
Mode d'apprentissage de la langue	3 « natale » 2 « héritée » 1 « acquise »	5 « natale »	6 « natale » 4 « acquise »

TABLE 2 : Statistiques sur les annotateurs par campagne

Le mode d'apprentissage de la langue visée diffère entre les annotatrices et annotateurs (cf. Table 2). Pour beaucoup, c'est une langue « natale », apprise dès la naissance, pour d'autres, une langue « héritée », correspondant aussi à un apprentissage par l'environnement familial mais plus tardif, et pour d'autres encore, une langue « acquise » par la formation. Cette variété est positive pour notre champ d'application mais peut engendrer un biais de compétence à intégrer dans l'analyse.

Afin d'évaluer comparativement l'impact des outils d'assistance, nous avons conçu trois types de groupe d'annotateurs : un groupe témoin « Sans assistance » dont les annotateurs n'ont jamais accès aux outils d'assistance, un groupe « Avec assistance » avec les outils d'assistance constamment à disposition, et a un type de groupe entre-deux, c'est le groupe « Alternance », qui passe d'une session d'annotation avec les outils d'annotation à une session sans, ou inversement. Chaque type de groupe a été représenté par un à deux annotateurs.

3.5 Les techniques d'assistance étudiées

Nous avons étudié une aide à la sélection de catégorie, qui consiste en la suggestion d'une étiquette par deux systèmes non supervisés : un système standard basé sur une technique « Zero-shot » (Pourpanah & Abdar, 2022), et un système innovant qui associe le précédent à la technique BERTopic (Grootendorst, 2022). Nous avons aussi mis en œuvre une aide à la sélection de documents, regroupant des outils standards permettant de filtrer les documents (mots-clés, thèmes, …). Enfin, une assistance d'aide à la lecture est apparue comme un besoin récurrent durant les deux premières campagnes, et a été étudiée pour la dernière campagne, sur l'arabe. Elle consiste en la mise en exergue des phrases discriminantes et l'identification d'entités nommées dans les articles.

Les campagnes sur l'hindi et sur le mandarin incluent un seul type d'assistance, que nous nommons « assistance A » et qui correspond à la combinaison de l'aide à la sélection de catégorie et de l'aide à la sélection de documents. La campagne sur l'arabe a la particularité d'inclure un second type d'assistance, l'« assistance B », qui est l'aide à la lecture.

3.6 Le déroulement

La première séquence de nos campagnes d'annotation est dédiée à la présentation des enjeux, du déroulement et du guide d'annotation. Nous recommandons notamment aux annotatrices et annotateurs de privilégier la qualité à la quantité. Les deux autres séquences comprennent des sessions d'annotation, définies par une durée et un sous-corpus, dans différents formats que nous préciserons. Adoptant une organisation agile (cf. section 2.1), à la fin de chaque session d'annotation, nous demandons aux annotatrices et annotateurs de remplir un formulaire de retour d'expérience, puis nous les faisons participer à une réunion de réconciliation qui vise à résoudre les cas d'incertitude et à obtenir une annotation commune lorsqu'il y a un désaccord. Ces sessions duraient de 15 à 40 minutes selon le temps disponible et permettaient de traiter environ 5 à 20 documents. La fin du cycle est marquée par la révision éventuelle du guide d'annotation.

La première séquence d'annotation est dédiée à la formation et à l'évaluation initiale des annotatrices et annotateurs, qui sert à quantifier un éventuel biais de compétence. Cela se déroule pour toutes et tous dans chacun des modes d'annotation étudiés par campagne : sans assistance, avec l'assistance A pour les trois campagnes et, pour la campagne sur l'arabe, avec l'assistance B. La formation consiste en la prise en main de la plateforme d'annotation par des sessions d'entraînement de 10 minutes et une vingtaine de documents proposés ; exceptionnellement, les annotatrices et annotateurs était autorisés à échanger entre eux pendant l'annotation. Les sessions d'évaluation initiale durent 45 minutes avec environ 130 documents proposés. La seconde séquence d'annotation, comprend quatre sessions de 50 minutes, avec environ 140 documents proposés, embarquant différents modes par la répartition des annotateurs dans les groupes tels que décrits en section 3.4.

4 Analyse

4.1 Méthodologie

L'analyse d'impact des techniques d'assistance a pour objectif de déterminer si celles-ci améliorent, détériorent ou n'ont pas d'effet notable sur le travail des annotatrices et annotateurs à travers les trois campagnes. Elle porte sur deux critères de performance : la qualité et la vitesse d'annotation.

Dans un premier temps, nous quantifions l'éventuelle différence de compétence des annotatrices et annotateurs entre les modes. Pour ce faire, nous calculons la différence entre les résultats des évaluations initiales reliés aux modes suivant la proportion de représentation des groupes lors des sessions en embarquant plusieurs. Par exemple, les annotatrices et annotateurs « Avec assistance A » lors des sessions embarquant différents modes affichent une moyenne d'évaluations initiales à 0,85 de F-Mesure, tandis que celles et ceux ayant pratiqué le mode « Sans assistance » produisent une F-Mesure moyenne à 0,82. Le biais de compétence en faveur du mode « Avec A » est égal à 0,85 - 0,82, soit 3 points de F-Mesure, dont nous tiendrons compte dans le calcul d'impact de l'assistance A.

$$\text{Impact(Avec)} = \text{Résultat(Avec)} - \text{Résuiltat(Sans)} - \text{Biais(Avec)}$$

ÉQUATION 1 : Formule de calcul d'impact d'une assistance

Dans un second temps, nous comparons les moyennes des résultats de vitesse et de qualité selon les modes à l'échelle des sessions en embarquant plusieurs, avec le mode « Sans assistance » servant de mode témoin. Nous intégrons à cette comparaison le biais de compétence décrit précédemment afin que les différences mesurées reflètent essentiellement l'impact des outils d'assistance (cf. Équation 1).

Pour évaluer la qualité des annotations, nous les comparons à une « référence », c'est-à-dire un ensemble d'annotations que nous considérons comme suivant précisément le guide d'annotation. Pour constituer cette référence, nous utilisons d'une part les annotations issues de la réconciliation et d'autres part celles correspondant à un certain niveau d'accord. Pour l'hindi et le mandarin, l'accord minium requis est d'au moins trois annotatrices ou annotateurs, tout le monde s'étant exprimé (environ 400 documents). Pour l'arabe, dont la campagne a davantage d'annotatrices et annotateurs (cf. section 3.4), le niveau minimum d'accord est d'au moins quatre annotatrices ou annotateurs et au maximum une divergence (environ 120 documents).

Nous utilisons la formule de la F-mesure par annotatrice ou annotateur G sur un sous-ensemble de documents qui est l'intersection de l'ensemble des documents de la référence et de l'ensemble des documents annotés par G. Autrement dit, nous n'évaluons la qualité qu'à partir des documents sur lesquels il y a une annotation de référence et une annotation de G. En ce qui concerne la vitesse

d'annotation, nous utilisons le volume de documents annotés, en pourcentage du volume de documents proposés par session (cf. section 3.6), et la durée d'annotation.

4.2 Résultats

Métrique	F-mesure	Volume (%)	Durée
Assistance A	+3,3pts	+2,43pts	0min
Assistance B	+2pts	-1,5pt	-2min

TABLE 3 : Statistiques sur les annotateurs par campagne

L'analyse des sessions embarquant différents modes à travers les trois campagnes semble montrer que l'assistance A aide à gagner en qualité d'annotation, avec une moyenne à 3,3 points supplémentaires de F-Mesure. L'impact de l'assistance A sur la vitesse d'annotation semble globalement négligeable, avec cependant 2,43 points d'impact positif en moyenne sur le pourcentage de volume d'annotation. Les retours d'expérience disent que la suggestion de catégorie, par un niveau de qualité variable, a parfois eu tendance à créer davantage de remises en question voire de confusion chez les annotateurs, induisant une vitesse d'annotation amoindrie.

L'assistance B, qui ne concerne quant à elle que la campagne sur l'arabe, semble avoir produit un gain en qualité évalué à 2 points de F-Mesure. Concernant la vitesse d'annotation, nous avons relevé globalement un impact négligeable, avec une perte en moyenne de 1,3 point en pourcentage de volume annoté mais 2 minutes en moins de durée d'annotation. D'après les retours d'expérience, l'assistance B a aidé à prendre connaissance du contenu des documents plus rapidement, par la mise en exergue des passages clés ainsi que des entités nommées. Cependant, la pré-catégorisation des passages était parfois fausse, engendrant un temps de réflexion supplémentaire.

5 Conclusion et perspectives

Grâce à ce partenariat, nous avons pu étudier avec succès l'apport de certaines assistances durant la phase d'amorçage d'une campagne d'annotation, et constaté des gains mesurables lorsqu'elles sont de bonnes qualités. Si la démarche semble donc validée, il n'en reste pas moins que la qualité des systèmes est une piste d'amélioration déterminante pour fournir une assistance efficace et aboutir à des outils qui soient présentés à des clients de Kairntech.

Dans la suite du projet Vital, nous allons étudier la consolidation des données produites, avec des annotations disponibles ouvrant la voie à des assistances supervisées ou semi-supervisée. Nous étudierons aussi la clôture d'une campagne d'annotation, qui consiste à finaliser le jeu de données en revisitant ses points faibles et en déterminant le point d'arrêt optimal.

Références

ABBAS, M., & SMAILI, K. (2005). Comparison of topic identification methods for arabic language. In Proceedings of International Conference on Recent Advances in Natural Language Processing, RANLP (pp. 14-17).

ALEX, B., NISSIM, M., & GROVER, C. (2006, May). The Impact of Annotation on the Performance of Protein Tagging in Biomedical Text. In *LREC* (pp. 595-600).

ARTSTEIN, R., & POESIO, M. (2008). Inter-coder agreement for computational linguistics. *Computational linguistics*, *34*(4), 555-596. DOI : 10.1162/coli.07-034-R2

BAYERL, P. S., & PAUL, K. I. (2011). What determines inter-coder agreement in manual annotations? A meta-analytic investigation. *Computational Linguistics*, *37*(4), 699-725. DOI : 10.1162/COLI_a_00074

EGGER, R., & YU, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. Frontiers in sociology, 7, 886498. DOI : 10.3389/fsoc.2022.886498

FAZAKIS, N., KANAS, V. G., ARIDAS, C. K., KARLOS, S., & KOTSIANTIS, S. (2019). Combination of active learning and semi-supervised learning under a self-training scheme. *Entropy*, 21(10), 988. DOI : 10.3390/e21100988

FORT, K. (2012). Les ressources annotées, un enjeu pour l'analyse de contenu: vers une méthodologie de l'annotation manuelle de corpus (Doctoral dissertation, Université Paris-Nord-Paris XIII). HAL : hal-00797760.

GROOTENDORST, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. DOI : 10.48550/arXiv.2203.05794

KRIPPENDORFF, K. (2004). Measuring the reliability of qualitative text analysis data. *Quality and quantity*, *38*, 787-800. DOI : 10.1007/s11135-004-8107-7

MATHET, Y. & WIDLÖCHER, A. (2016). Évaluation des annotations : ses principes et ses pièges. *Revue TAL*, 57-2, 73-98. HAL : hal-01712282

POURPANAH, F., ABDAR, M., LUO, Y., *et al* (2022). A review of generalized zero-shot learning methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. DOI : 10.1109/TPAMI.2022.3191696

RIFKIN, R. (2008). Multiclass classification. *Lecture Notes, Spring08. MIT, USA*, 59.

SCHNOBER, C., & GUREVYCH, I. (2015). Combining topic models for corpus exploration: applying LDA for complex corpus research tasks in a digital humanities project. In *Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications* (pp. 11-20). DOI : 10.1145/2809936.2809939

WINATA, G. I., MADOTTO, A., LIN, Z., LIU, R., YOSINSKI, J., & FUNG, P. (2021). Language models are few-shot multilingual learners. DOI : 10.48550/arXiv.2109.07684

WU, S., & DREDZE, M. (2020). Are all languages created equal in multilingual BERT? DOI : 10.48550/arXiv.2005.09093

Extraction de règles de grammaire à partir de treebanks : développement d'un outil et premiers résultats

Santiago Herrera[1] (en collaboration avec Sylvain Kahane[1] et Bruno Guillaume[2])
(1) MoDyCo, Université Paris Nanterre, France
(2) Loria, Nancy, France
santiago.herrerayanez@parisnanterre.fr, sylvain@kahane.fr,
bruno.guillaume@loria.fr

RÉSUMÉ

Ce travail présente une méthode et un outil d'extraction et d'exploration automatique de motifs statistiquement significatifs, de potentielles règles de grammaire, à partir de corpus arborés.

ABSTRACT

Grammar rules extraction from treebanks : a tool and some first results

This work presents a method and a tool for automatic extraction and exploration of statistically significant patterns and potential grammar rules from treebanks.

MOTS-CLÉS : Extraction de grammaire, règles de grammaire, treebank.

KEYWORDS: Grammar extraction, grammar rules, treebank.

1 Introduction

Construire la grammaire d'une langue et repérer l'ensemble de ses règles est une tâche aussi fondamentale pour l'étude de la langue et pour le développement d'autres ressources langagières que coûteuse. Plusieurs domaines, notamment la typologie linguistique (Dryer & Haspelmath, 2013), la linguistique formelle (Bender *et al.*, 2014; Howell *et al.*, 2017) et le traitement automatique des langues (Ponti *et al.*, 2019; Chaudhary *et al.*, 2020), cherchent depuis longtemps à systématiser les contraintes des langues, de façon plus ou moins automatique, afin de rendre compte de leurs propriétés structurelles. La plupart de ces approches s'appuient sur des corpus annotées à travers lesquels il est possible d'exploiter les propriétés statistiques du langage et se concentrent soit sur des règles générales, soit sur l'ensemble des structures possibles d'une langue. Pour notre part, nous cherchons à rendre compte des règles dans un espace local, en les considérant comme une série de contraintes qui déclenchent une sur-représentativité d'un motif donné spécifique dans un contexte précis.

Ce travail présente une méthode et une première version d'un outil d'extraction et d'exploration automatique de motifs statistiquement significatifs et de potentielles règles de grammaire, à partir de corpus annotés en syntaxe ou treebanks en dépendance. On vise plus précisément à l'extraction de règles interprétables et pondérables selon leurs propriétés quantitatives et statistiques. Inspirés par des travaux de la lexicométrie (Lafon, 1980) et de la linguistique du corpus (Evert, 2005; Pecina, 2010), on tire profit des tests et des mesures statistiques qui permettent le repérage de motifs, leur classement et la comparaison des motifs significatifs entre différents corpus. Ce système d'extraction

Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain *(LIFT)* et Traitement Automatique des Langues *(TAL)*,
Marseille, France, 14 au 15 novembre 2022 , pages 93–98,

est implémenté à travers un outil [1] accessible qui combine ces méthodes statistiques avec l'utilisation de GREW-MATCH (Guillaume, 2021), afin d'interroger et d'explorer les arbres syntaxiques.

Le travail présenté ici est issu de Herrera (2022) et il a été réalisé dans le cadre du projet ANR Autogramm (Kahane, 2021), dont un des objectifs est de construire conjointement des treebanks et des grammaires pour des langues peu décrites. Notre outil cherche justement à faciliter le développement de grammaires basées sur (peu) des données avec le but d'aider le linguiste dans la description d'une langue.

2 Hypothèses

Une règle de grammaire est une contrainte d'une langue qui montre une régularité relative dans le système de cette langue. Dans un corpus, cette régularité se traduit par la répétition élevée et non aléatoire d'un motif, lequel est représenté dans une proportion inattendue par rapport à un sous-ensemble pertinent de motifs. Une règle explicite aussi les conditions qui, dans un contexte donné, déclenchent un motif particulier de façon statistiquement élevée.

Nous définissons donc une règle de grammaire à partir de trois éléments :

$$M \Rightarrow M(X) \,|\, C$$

(a) Définition générale d'une règle de grammaire.

$$M1 \Rightarrow M1\&M2 \,|\, M3$$

(b) Définition opérationnelle à partir de motifs d'une règle de grammaire.

FIGURE 1 – Formalisation d'une règle de grammaire.

Partant d'un motif M donné ou d'un espace de recherche spécifique, on cherche les conditions C qui favorisent de façon statistiquement significative les occurrences d'un phénomène linguistique X dans M. Autrement dit, on veut extraire et mettre en évidence quelles sont les variables ou motifs C qui montrent une dépendance statistique, positive en termes d'occurrences, avec X dans M. Cette formalisation s'inspire des règles de correspondance de la Théorie Sens-Texte (Mel'čuk, 1988), bien que l'on cherche à représenter des relations de dépendance entre variables et non pas une correspondance entre différents niveaux de représentation. Les règles sur lesquelles nous travaillons sont plus similaires aux règles de bonne formation ou de linéarisation. L'utilisation de la notion de contrainte trouve également un écho dans les travaux de Grammaires des Propriétés (Blache, 2001), où l'accent est mis sur les contraintes ou *propriétés* qui règlent la combinaison des unités et la formation des phrases. Néanmoins, nos règles sont motivées et limitées par l'information contenue dans le corpus.

Nous supposons que les conditions ou prédicteurs qu'on peut prendre en compte sont présents dans les contextes immédiats possibles du motif M considéré. Dans un arbre de dépendance, c'est la tête syntaxique d'une unité qui détermine dans la plupart de cas les propriétés linguistiques de cette unité. Mais toute autre relation, co-dépendance ou information linguistique encodée dans la phrase ou dans le treebank sont exploitables. Pour extraire de règles, nous élaborons une hypothèse de « localité » laquelle on suppose que les prédicteurs se trouvent dans le contexte immédiat du motif M. Notre espace de recherche (voir Figure 2) se limite donc aux différents traits des nœuds et des dépendances de M, ainsi que des nœuds directement connectés à M.

1. Voir https://github.com/santiagohy/grammar-rules-extraction

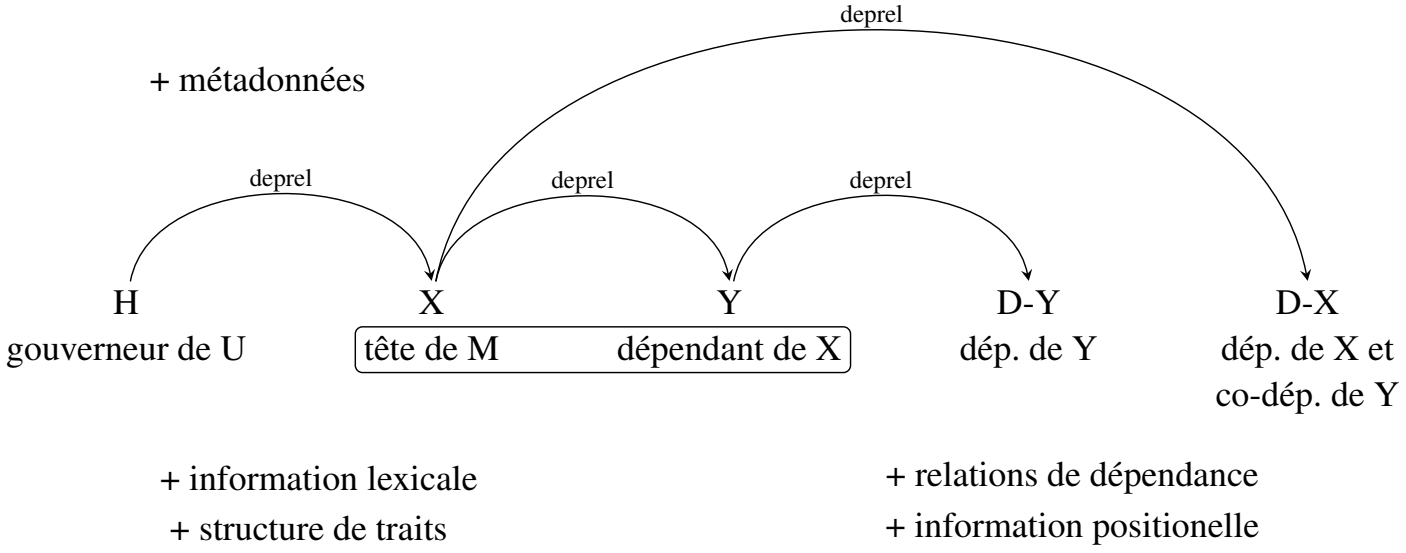


FIGURE 2 – Espace de recherche où X et Y forment, à titre d'exemple, le motif M. La structure de traits inclue les informations encodées dans les nœuds, notamment les traits morphologiques, les lemmes, etc.

3 Méthode d'extraction

Afin de rendre opérationnelle nos hypothèses, on peut formaliser l'extraction d'une règle à partir de trois motifs (Kahane, 2021; Kahane *et al.*, 2021). On cherche les motifs M3 qui, étant donné un motif M1, favorisent significativement les occurrences de M1&M2. Par exemple (voir Table 1), dans l'ensemble des noms modifiés par des adjectifs (M1), le fait que l'adjectif soit un numéral ordinal (M3) favorise significativement l'antéposition de l'adjectif (M1&M2) [2].

	Formalisation	**Exemple**
M1	Espace de recherche de départ	X->Y ; X[upos=NOUN] ; Y[upos=ADJ]
M2	Variable dépendante	Y << X
M3	Variable(s) explicative(s)	Y[NumType=Ord]

TABLE 1 – Formalisation de l'extraction avec l'exemple simple de l'antéposition de l'adjectif en français. On utilise le langage de requête GREW et les étiquettes morphosyntaxiques et syntaxiques UD/SUD. Dans l'exemple, X->Y indique une relation de dépendance orientée entre les nœuds X et Y, et Y << X signale la position relative entre ces deux nœuds dans la phrase, Y étant avant X. Les traits de chaque nœud, avec leurs valeurs, sont explicités entre crochets.

Pour déterminer si un motif M3 est statistiquement significatif, on calcule la probabilité d'obtenir la distribution observée, ou une distribution plus extrême, sous l'hypothèse nulle selon laquelle les motifs M2 et M3, dans le contexte M1, seraient indépendants. Si la probabilité est inférieure à la valeur critique fixée (p-value < 0.01), on rejette l'hypothèse nulle et on considère le motif significatif. Le logarithme de cette probabilité est utilisé comme valeur de significativité.

2. Pour un étude plus approfondie sur la position de l'adjectif et l'ordre de mots en français, voir Thuilier (2012).

Nous choisissons d'utiliser le test exact de Fisher pour calculer la significativité d'un motif à partir de ses occurrences car il nous permet de travailler avec des échantillons de petite taille et il est donc plus adapté au travail avec des langues pour lesquelles il y a peu de ressources (voir Lafon (1980) pour une utilisation de la méthode en lexicométrie). La tâche que nous réalisons se différencie des travaux connexes qui cherchent à évaluer la contribution de la combinaison linéaire d'un ensemble de variables binaires et numérique, dans la prédiction d'un phénomène linguistique (Bresnan *et al.*, 2007; Thuilier, 2012). Nous cherchons, en revanche, à extraire les meilleures prédicteurs, facilement interprétables, à partir d'un vaste ensemble composé, dans la plupart des cas, de variables catégorielles non binaires.

Il s'agit d'une méthode non-déterministe, où les motifs extraits peuvent être des motifs non disjoints. Nous privilégions donc le classement de ces motifs. C'est la raison pour laquelle nous utilisons aussi d'autres mesures complémentaires pour décrire et classer les motifs extraits, notamment la taille d'effet et les proportions des occurrences trouvées sur la totalité des motifs M1&M2 et sur la totalité des motifs M1&M3. La première de ces deux proportions peut s'interpréter comme la couverture de la règle sur le motif et le phénomène linguistique que l'on veut expliquer. La deuxième représente la précision de la règle ou combien de motifs M3 sont effectivement concernés par elle.

Motif	Significativité	OR	Proportion sur M1&M2	Proportion sur M1&M3
Y[NumType=Mod]	89	3.77	18.63%	92.22%

TABLE 2 – Résultat pour l'exemple sur le treebank SUD_French-Sequoia version 2.10.On prend le logarithme arrondi de la p-valeur comme valeur de la significativité. La valeur du OR est plus précisément le logarithme du rapport des chances ou, en anglais, *Odds Ratio*.

En reprenant l'exemple de la place de l'adjectif épithète, sur l'ensemble des traits de l'adjectif (nœud Y), que l'adjectif soit un numéral ordinal est très significatif du fait qu'il se place avant le nom modifié. On obtient donc un nombre d'adjectifs placés avant le nom très inattendue. Plus précisément, avoir obtenu 89 comme valeur de significativité signifie qu'il existe une probabilité très basse, d'autour de 10^{-89}, d'observer un nombre égale ou supérieur à la valeur observé de numéraux ordinaux placés avant leur nom, sous l'hypothèse d'indépendance. Ce motif significatif conforme, donc, une règle de linéarisation de l'adjectif par rapport au nom, qui exprime que les adjectifs dépendants d'un nom s'inversent très largement si ce sont des numéraux ordinaux. Cette règle explique que 18.63% des adjectifs antéposés sont des ordinaux et que plus de 92% des ordinaux se placent avant le nom qu'ils modifient.

4 Résultats et perspectives

La méthode nous permet d'obtenir les résultats que nous attendions. La plupart des règles construites sont des règles de grammaires (règles d'ordre, d'accord, de régime, etc.), bien qu'on extrait aussi des propriétés du corpus et des séries non pertinentes de motifs, d'un point de vue informatif. Ce dernier groupe de résultats est généralement constitué de sous-motifs d'autres motifs plus significatifs ou de motifs faisant partie d'une règle qui ne peut pas être expliquée par les informations linguistiques disponibles dans le corpus.

L'outil qui a été développé permet, d'autre part, d'interroger les règles d'un corpus arborés dans deux directions : de façon ascendante ou descendante. Dans le premier cas, il est possible de l'utiliser pour

explorer un treebank et pour découvrir les règles qui émergent d'un corpus précis, quelle que soit sa taille. Cela permet de mettre en parallèle l'annotation d'un corpus et la construction d'une grammaire, un des objectifs du projet ANR Autogramm, sachant que ces deux tâches peuvent mutuellement se compléter.

Dans le deuxième cas, rien n'empêche d'utiliser cet outil pour valider les règles d'une théorie linguistique précise de façon quantitative et statistique à partir d'un corpus donné. Ainsi, une règle théorique composée de conditions ou de contraintes, comme on peut en trouver dans la Théorie Sens-Texte ou dans le cadre des Grammaires de Propriétés, sera décrite selon sa significativité statistique dans le corpus et à partir des autres valeurs expliquées. Cela favorise le dialogue entre les théories linguistiques et les données empiriques, mettant dans la mesure du possible en relief la possible distance entre les deux et en hiérarchisant les règles selon leur importance dans le corpus. Le travail à partir de corpus de différente nature représente aussi l'occasion de valider ou de mettre à jour certaines affirmations théoriques en fonction du type ou de la variété de la langue étudiée.

L'extraction de règles, comme nous l'avons dit, dépend de l'information contenue dans l'espace de recherche et dans le corpus arborés. Nous travaillons pour l'instant avec les treebanks UD/SUD tels qu'ils se présentent, ce qui permet essentiellement d'extraire des règles syntaxiques de surface (ordre des mots et accord). Une prochaine étape consistera à travailler sur l'enrichissement automatique de treebanks, notamment sur le raffinage de certains traits pour obtenir des traits plus élémentaires et, nous espérons, plus informatifs.

Au moment de l'écriture de ce document, nous explorons d'autres méthodes statistiques et d'autres mesures d'association qui nous permettront d'extraire des motifs significatifs de façon plus efficace et d'avoir un outil plus performant. En parallèle au développement de l'outil, nous travaillions sur des méthodes d'évaluation quantitatives, notamment l'évaluation de la couverture de l'ensemble des règles extraites sur un treebank.

Remerciements

Les auteurs remercient les relecteurs pour leurs commentaires. Ce travail a été réalisé dans le cadre du projet ANR Autogramm (ANR-21-CE38-0017).

Références

BENDER E. M., CROWGEY J., GOODMAN M. W. & XIA F. (2014). Learning grammar specifications from IGT : A case study of chintang. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, p. 43–53, Baltimore, Maryland, USA : Association for Computational Linguistics. DOI : 10.3115/v1/W14-2206.

BLACHE P. (2001). *Les grammaires de propriétés : Des contraintes pour le Traitement Automatique des Langues Naturelles*. Hermés.

BRESNAN J., CUENI A., NIKITINA T. & BAAYEN H. (2007). Predicting the dative alternation. *Cognitive Foundations of Interpretation*, p. 69–94.

CHAUDHARY A., ANASTASOPOULOS A., PRATAPA A., MORTENSEN D. R., SHEIKH Z., TSVETKOV Y. & NEUBIG G. (2020). Automatic extraction of rules governing morphological agree-

ment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 5212–5236, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.422.

DRYER M. S. & HASPELMATH M., Éds. (2013). *WALS Online*. Leipzig : Max Planck Institute for Evolutionary Anthropology.

EVERT S. (2005). *The Statistics of Word Cooccurrences : Word Pairs and Collocations*. Thèse de doctorat, University of Stuttgart.

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*, p. 168–175, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.eacl-demos.21.

HERRERA S. (2022). Extraction automatique de règles de grammaire à partir de treebanks. Mémoire de master, Master pluriTAL, Université Sorbonne Nouvelle, Université Paris Nanterre et l'INALCO.

HOWELL K., BENDER E. M., LOCKWOOD M., XIA F. & ZAMARAEVA O. (2017). Inferring case systems from IGT : Enriching the enrichment. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, p. 67–75, Honolulu : Association for Computational Linguistics. DOI : 10.18653/v1/W17-0110.

KAHANE S. (2021). Autogramm : Induction of descriptive grammars from annotated corpora. *Soumission à l'Agence National de la Recherche*.

KAHANE S., GUILLAUME B., GERDES K., CARON B. & LOISEAU S. (2021). Le projet ANR Autogramm et l'extraction automatique de grammaires. llustration par la négation. In *3e Journées GdR LIFT, Linguistique Informatique Formelle et de Terrain*, Grenoble, France.

LAFON P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots. Les langages du politique*, **1**(1), 127–165. DOI : 10.3406/mots.1980.1008.

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

PECINA P. (2010). Lexical association measures and collocation extraction. *Language Resources and Evaluation*, **44**(1/2), 137–158.

PONTI E. M., O'HORAN H., BERZAK Y., VULIĆ I., REICHART R., POIBEAU T., SHUTOVA E. & KORHONEN A. (2019). Modeling language variation and universals : A survey on typological linguistics for natural language processing. *Computational Linguistics*, **45**(3), 559–601. DOI : 10.1162/coli_a_00357.

THUILIER J. (2012). *Contraintes préférentielles et ordre des mots en français*. Thèses, Université Paris-Diderot - Paris VII.

Extraction et analyse de concepts médicaux dans un corpus de spécialité en orthophonie

Tiphaine Le Clercq de Lannoy[1] Romaric Besançon[1] Olivier Ferret[1]
Julien Tourille[1] Frédérique Brin-Henry[2] Bianca Vieru[1]
(1) Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
(2) CH Bar le Duc, ATILF UMR7118 CNRS-Université de Lorraine, 54063 Nancy, France
{prénom.nom}@cea.fr, fhenry@atilf.fr

1 Introduction

L'émergence de gros modèles de langue pré-entraînés tels que BERT (Devlin *et al.*, 2019) a développé la définition et l'application de stratégies d'apprentissage par transfert (*transfer learning*), en particulier par le biais de la notion d'affinage (*fine-tuning*). Bien que ce développement facilite l'apprentissage de modèles pour des domaines spécialisés à partir de modèles plus généraux, cet apprentissage souffre toujours de l'absence de données annotées en quantités suffisantes. Dans cet article, nous nous focalisons plus spécifiquement sur le domaine de la santé et sur la tâche de reconnaissance d'entités nommées en français. Nous explorons plus précisément deux voies pour faciliter l'adaptation aux domaines spécialisés. La première reprend l'idée, explorée initialement par Gururangan *et al.* (2020), qu'utiliser un corpus non annoté du domaine cible et l'utiliser afin de poursuivre l'entraînement d'un modèle pré-entraîné sur sa tâche de modélisation du langage permet de spécialiser ce modèle pour ce domaine et d'améliorer les résultats de l'affinage sur la tâche finale visée. Cette approche a été appliquée en particulier par Copara *et al.* (2020) pour la reconnaissance d'entités nommées médicales en français.

La seconde voie exploite quant à elle les connaissances existant pour le domaine cible, connaissances qui sont particulièrement riches dans le cas du domaine médical et de la santé. Plus précisément, parmi les nombreux travaux réalisés pour utiliser conjointement les modèles de langue neuronaux et des connaissances données a priori (Yin *et al.*, 2022; Wei *et al.*, 2021; Yang *et al.*, 2022), se distinguent les approches que l'on peut qualifier de précoces, visant à injecter les connaissances directement au sein des modèles, soit lors de leur construction, soit a posteriori, des approches dites tardives dans lesquelles modèles de langue et connaissances sont fusionnés au niveau des résultats liés à la tâche. Nous nous situons dans cette seconde perspective en nous distinguant néanmoins des approches de type auto-apprentissage (Gao *et al.*, 2021) dans lesquelles les connaissances sont utilisées pour réaliser une forme d'augmentation de données.

De plus, nous appliquons les techniques étudiées à un corpus d'orthophonie, OrthoCorpus (2019), afin d'analyser les extractions d'entités nommées sur des cas concrets, du point de vue de l'intérêt clinique de la démarche et de sa faisabilité pour les experts du domaine. D'un point de vue disciplinaire, cela permet en effet de questionner le classement conceptuel en santé dans un sous-domaine spécifique au carrefour des sciences biomédicales et des sciences humaines et sociales. L'examen des formes et du statut des candidats-termes [1] nous renseigne sur la langue de spécialité (L'Homme, 2011).

1. ISO 1087-1:2000 : un terme est une désignation représentant un concept général d'un domaine spécifique ou d'un sujet.

Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain *(LIFT) et* Traitement Automatique des Langues *(TAL),*
Marseille, France, 14 au 15 novembre 2022 , pages 99–108,

Plus précisément, au travers des contributions de cet article, nous montrons, pour la reconnaissance d'entités nommées dans le domaine de la santé, que :

— l'utilisation de corpus spécialisés pour l'adaptation de modèles de langue pré-entraînés peut être intéressante, même pour des corpus que l'on peut qualifier de petits vis-à-vis des expérimentations de Gururangan *et al.* (2020) ;
— différents modèles neuronaux et une approche à base de connaissances présentent des profils complémentaires qu'une combinaison tardive permet de valoriser.

2 Approche

Pour entraîner, malgré des données annotées en quantité limitée, un modèle de langue spécialisé pour la reconnaissance d'entités nommées médicales en français, nous nous appuyons principalement sur deux éléments : l'exploitation de connaissances structurées dans le domaine de la santé, sous forme principalement de thésaurus (cf. section 2.1), et l'adaptation d'un modèle de langue au domaine de la santé (cf. section 2.2). Comme les deux approches présentées précédemment reposent sur des techniques très différentes, les résultats obtenus par chacune d'entre elles peuvent se compléter efficacement (cf. section 2.3).

2.1 Exploitation de connaissances

Pour notre approche à base de connaissances, nous avons retenu une méthode comparable à QuickUMLS (Soldaini & Goharian, 2016), fondée sur la projection dans le corpus cible d'une terminologie de référence, structurée selon les types d'entités visés. Une dimension essentielle de cette approche est donc la constitution de cette terminologie, structurée dans notre cas selon les dix groupes de types sémantiques de l'UMLS [2] (Unified Medical Language System (Lindberg *et al.*, 1993)) retenus pour annoter le corpus QUAERO (Névéol *et al.*, 2014), utilisé dans nos évaluations. L'orthophonie étant une profession de santé, l'application de ces mêmes groupes de types sémantiques permet la comparaison des résultats dans ce domaine de spécialité avec ceux obtenus plus généralement dans le domaine médical. Il s'agit plus précisément des groupes : *Anatomy*, *Chemicals & Drugs*, *Devices*, *Disorders*, *Geographic Areas*, *Living Beings*, *Objects*, *Phenomena*, *Physiology et Procedures*.

Cette terminologie est issue de plusieurs sources, à commencer bien entendu par l'UMLS lui-même puisque ce dernier contient un ensemble significatif de termes en français pour les dix groupes considérés, en particulier issus des terminologies MeSH, MedDRA et LOINC. Nous utilisons également les ressources constituées par Embarek & Ferret (2008) pour les types d'entités *Anatomy*, *Chemicals & Drugs*, *Disorders* et *Procedures*. Nous avons par ailleurs exploité les données du site de la base de données publique des médicaments (ANSM, 2021), qui référence tous les médicaments en vente sur le marché français ou en arrêt de commercialisation depuis moins de trois ans. Cette base nous a ainsi permis d'étoffer le type *Chemicals & Drugs* avec des médicaments et le type *Disorders* avec certaines pathologies. Pour finir, le site PasseportSante [3], recommandé par RESEAU CHU [4], nous a permis d'obtenir des termes plus grand public pour les types *Anatomy*, *Disorders* et *Procedures*.

Pour identifier dans les textes les types d'entités considérés à partir de ces ressources terminologiques,

2. https://www.nlm.nih.gov/research/umls/index.html
3. https://www.passeportsante.net/
4. https://www.reseau-chu.org/

nous avons défini et implémenté l'outil QuickMatching, fondé sur l'algorithme SimString (Okazaki & Tsujii, 2010), à l'instar de QuickUMLS. Cet outil calcule la similarité entre des termes de référence et les mots des textes sur la base d'un découpage en n-grammes. Cette mesure de similarité permet d'apparier un terme de référence avec un mot du texte en faisant abstraction de différences minimes, comme celles résultant de variations morphologiques mineures ou de fautes de frappe.

2.2 Adaptation de modèles de langue

Nous traitons la tâche d'identification des types d'entités médicales visés comme une tâche d'annotation de séquences au format BIO. Pour cela, nous reprenons l'architecture de Devlin *et al.* (2019) pour la tâche de reconnaissance d'entités nommées mais en utilisant CamemBERT (Martin *et al.*, 2020) comme modèle de langue initial. Pour l'adaptation de ce dernier au domaine médical, nous poursuivons sa tâche de modélisation du langage sur un corpus de textes du domaine de la santé. Nous présentons dans le tableau 1 les corpus utilisés pour cette adaptation, sélectionnés à la fois pour la possibilité de les obtenir facilement et l'absence de difficulté vis-à-vis de la problématique des données personnelles. Pour étudier à la fois l'influence de la taille des corpus et de leur nature sur une telle adaptation, nous avons entraîné un modèle spécifique pour chaque corpus ainsi qu'un modèle s'appuyant sur l'ensemble de ces corpus, ce qui représente un peu plus de 136 millions de mots.

Corpus	Description	Taille
OrthoCorpus (2019)	Articles de la revue spécialisée Rééducation Orthophonique (Brin-Henry, 2018)	6,7M
ISTEX	Articles de revues médicales indexées par ISTEX (Inist)	42,6M
EQueR	Articles scientifiques et de recommandations de bonne pratique médicale (CISMeF)	16,8M
PMC OA	Articles de revues médicales (PubMed Central Open Access)	3,8M
Cochrane	Résumés d'articles de l'organisation Cochrane	5,0M
EMA	Notices de l'Agence Européenne des Médicaments	21,2M
CRTT	Articles de revues, extraits de Science Direct	21,7M
E3C-Corpus	Résumés d'articles, articles de revues, cas cliniques	12,1M
Wikipédia	Articles Wikipédia dans le domaine médical	6,6M

TABLE 1 – Collections de textes du domaine de la santé utilisées. Les tailles sont exprimées en millions de mots

2.3 Combinaison des approches

L'exploitation de connaissances permet d'extraire des termes avec une bonne précision mais repose sur la qualité et la mise à jour de la terminologie de référence qui la sous-tend. L'utilisation des modèles de langue permet en revanche de généraliser l'annotation des termes vus durant l'entraînement mais nécessite des corpus annotés pour cette tâche. Pour combiner plusieurs approches, nous unissons les entités prédites par ces approches avec une gestion minimale des conflits au niveau des types. Plus précisément, si deux approches identifient une entité de même empan mais avec un type différent, la priorité est donnée au type trouvé par le modèle neuronal. Nous avons en effet constaté que donner la priorité à QuickMatching se traduisait par une dégradation des performances de l'ordre d'un point de F1-mesure.

Une version améliorée de cette combinaison d'approches, le vote, utilisée par Copara *et al.* (2020), consiste à utiliser plus de deux modèles et à procéder à un vote pour chaque entité prédite. Si deux modèles ou plus prédisent une entité, elle est alors conservée dans la version finale, ce qui

permet de réduire le bruit par rapport à la méthode précédente. Pour ce vote, nous considérons l'outil QuickMatching, le modèle CamemBERT pré-entraîné sur tous les corpus ainsi que le modèle Pyramid (Wang *et al.*, 2020), entraîné sur le corpus QUAERO et utilisé pour générer des entités imbriquées.

3 Expérimentations et résultats

3.1 Cadre expérimental

Données Nous évaluons les méthodes proposées sur le corpus QUAERO, annoté en entités médicales pour le français et utilisé dans le challenge CLEF eHealth 2016 (Névéol *et al.*, 2016). Ce corpus est composé de dix documents sur des médicaments issus de l'European Medicines Agency (EMEA) ainsi que de 2 498 titres d'articles de recherche disponibles dans la base de données de MEDLINE. Les types d'entités utilisés pour annoter le corpus correspondent aux dix groupes de types sémantiques de l'UMLS évoqués à la section 2.1. Il est à noter que cette annotation comporte des entités imbriquées, le nombre de niveaux d'imbrication pouvant aller jusqu'à quatre. Il n'y a pas de restrictions sur les types utilisés dans les entités imbriquées. Pour l'évaluation, nous considérons toutes les entités au niveau de la référence. En revanche, nos deux méthodes de base se comportent de façon différente : tandis que QuickMatching peut identifier des entités imbriquées, notre modèle neuronal est entraîné pour identifier seulement les entités de plus large extension, ce qui le désavantage nécessairement du point de vue du rappel. Compte tenu de la méthode de fusion, son résultat comporte les entités imbriquées issues de QuickMatching.

Entraînement des modèles Pour la manipulation des modèles pré-entraînés, nous nous sommes appuyés sur la bibliothèque Transformers de HuggingFace (Wolf *et al.*, 2020). Concernant l'adaptation du modèle de langue CamemBERT, nous avons appliqué la tâche de Masked Language Modeling (MLM) en masquant des mots entiers, à l'instar de Martin *et al.* (2020), et non les seuls WordPieces. Nous avons utilisé l'optimiseur Adam (Kingma & Ba, 2015), avec $\beta_1 = 0,9$ et $\beta_2 = 0,98$. Le taux d'apprentissage (*learning rate*) était égal à 2.10^{-5}. Pour chaque corpus, nous avons réalisé 15 époques (*epochs*) de MLM en sauvegardant le modèle à la fin de chaque époque et avons sélectionné la version du modèle obtenant les meilleurs résultats sur le jeu de validation du corpus QUAERO.

Concernant l'affinage sur la tâche de reconnaissance d'entités médicales, nous avons utilisé l'outil Optuna (Akiba *et al.*, 2019) pour la recherche des meilleures valeurs d'hyperparamètres en prenant en compte la taille des lots, le nombre d'époques, le taux d'apprentissage et le ratio d'échauffement (*warm-up*). Nous avons ainsi obtenu la combinaison : taille de lot $= 9$, taux d'apprentissage $= 8.10^{-5}$ et ratio d'échauffement $= 0,224$. L'ensemble du jeu d'entraînement de QUAERO (EMEA et MEDLINE) a été utilisé pour réaliser chaque affinage de modèle pré-entraîné.

L'entraînement du modèle Pyramid a été réalisé grâce au code fourni par les auteurs [5] en adoptant le modèle de base associé à une pyramide inverse, le tout d'une profondeur 9. Les plongements utilisés sont ceux de fastText (Bojanowski *et al.*, 2017) en français [6] et le taux d'apprentissage était de 0,01.

Métriques d'évaluation Pour évaluer les résultats des modèles, nous avons utilisé l'outil BRAT-Eval ehealth fourni avec le corpus QUAERO. Cet outil a été développé par (Verspoor *et al.*, 2013) et modifié par (Névéol *et al.*, 2014). Les métriques considérées sont la précision (P), le rappel (R) et la F1-mesure (F1). Ce sont des micro-mesures calculées en mode strict.

5. https ://github.com/LorrinWWW/Pyramid
6. https ://fasttext.cc/docs/en/crawl-vectors.html

Modèle	Test EMEA			Test MEDLINE		
	P	R	F1	P	R	F1
QuickMatching	62,6	**66,4**	64,4	60,9	**61,7**	**61,3**
Pyramid (Wang *et al.*, 2020)	67,0	58,5	62,4	59,0	53,7	56,2
CamemBERT	72,6 ± 1,5	60,1 ± 1,8	65,8 ± 1,5	62,4 ± 1,0	48,9 ± 0,9	54,8 ± 0,8
OrthoCorpus	73,4 ± 1,6	59,4 ± 2,0	65,7 ± 1,2	**63,1 ± 1,9**	47,6 ± 1,3	54,2 ± 0,6
PMC OA	71,5 ± 1,4	59,9 ± 1,4	65,2 ± 1,0	61,1 ± 0,7	48,1 ± 1,8	53,8 ± 0,9
Cochrane	72,1 ± 1,3	59,8 ± 1,5	65,3 ± 0,9	61,4 ± 0,8	47,8 ± 0,8	53,7 ± 0,5
EQueR	72,3 ± 1,1	60,5 ± 0,8	65,9 ± 0,4	61,9 ± 0,9	49,0 ± 1,1	54,7 ± 1,0
ISTEX	72,4 ± 1,4	60,0 ± 1,3	65,6 ± 1,2	63,0 ± 0,8	49,0 ± 0,8	55,1 ± 0,7
CRTT	73,4 ± 0,6	60,4 ± 1,7	66,3 ± 1,1	62,5 ± 1,2	48,6 ± 1,0	54,7 ± 0,6
E3C-Corpus	75,1 ± 1,3	61,8 ± 1,4	67,8 ± 1,0	61,7 ± 1,0	47,9 ± 0,7	53,9 ± 0,3
Wikipédia	72,9 ± 2,0	60,4 ± 1,7	66,1 ± 1,8	62,1 ± 1,5	48,6 ± 0,3	54,5 ± 0,6
EMA	**75,4 ± 0,8**	61,8 ± 1,1	**67,9 ± 0,9**	61,7 ± 2,1	47,8 ± 2,0	53,8 ± 2,0
Tous les corpus	73,4 ± 0,4	62,2 ± 0,6	67,4 ± 0,4	62,2 ± 1,3	49,7 ± 0,9	55,3 ± 1,0

TABLE 2 – Comparaison des références (QuickMatching, Pyramid et modèle CamemBERT entraîné sur QUAERO sans pré-entraînement et avec affinage) et des modèles pré-entraînés avec différents corpus. Les résultats sont donnés sous la forme de moyennes et écarts-types obtenus en utilisant cinq graines aléatoires

3.2 Résultats et discussion

Résultats des approches de base Le tableau 2 présente les résultats sur le test du corpus QUAERO de nos deux approches de base (lignes QuickMatching et CamemBERT) ainsi que des expériences d'adaptation de notre modèle neuronal par pré-entraînement sur différents corpus du domaine de la santé. Dans le cas de notre modèle neuronal, la condition de base correspond à un affinage à partir du modèle CamemBERT, sans pré-entraînement complémentaire. Les résultats pour tous les modèles neuronaux correspondent à des moyennes pour cinq graines aléatoires.

Nous constatons en premier lieu que le modèle présentant en moyenne les meilleurs résultats à la fois sur EMEA et sur MEDLINE est le modèle pré-entraîné sur tous les corpus. S'il n'est pas le meilleur sur le corpus EMEA, il est tout de même gratifié du meilleur rappel. Le meilleur modèle sur EMEA est obtenu en pré-entraînant CamemBERT avec EMA. Or, ces deux corpus proviennent tous deux de l'Agence Européenne des Médicaments et comportent donc de fortes similarités au niveau des types de documents ainsi que de leurs sujets. Ces résultats confirment ainsi deux tendances de fond : les performances en affinage bénéficient d'autant mieux des effets d'un pré-entraînement en MLM que celui-ci se fait sur un gros corpus. Néanmoins, la spécificité de ce corpus par rapport aux données de test a aussi son importance et un corpus plus petit mais plus spécialisé peut s'avérer plus efficace.

Concernant QuickMatching, nous remarquons que les résultats sont assez constants entre EMEA et MEDLINE, contrairement aux approches fondées sur CamemBERT. Comme les mêmes ressources sont utilisées pour obtenir les résultats sur les deux corpus, nous pouvons supposer qu'elles couvrent de manière équivalente les deux corpus.

Résultats des combinaisons La table 3 compare les différentes méthodes de combinaison de modèles décrites dans la section 2.3, l'union et le vote, toutes les deux réalisées sur les modèles CamemBERT pré-entraîné avec tous les corpus, QuickMatching et Pyramid.

Nous pouvons constater que, quelle que soit la combinaison utilisée, le rappel augmente significative-

Combinaison	Test EMEA			Test MEDLINE		
	P	R	F1	P	R	F1
Union	56,3 ± 0,2	**82,4 ± 0,4**	66,9 ± 0,2	51,6 ± 0,6	**76,6 ± 0,7**	61,7 ± 0,6
Vote	**80,3 ± 0,1**	65,6 ± 0,4	**72,2 ± 0,2**	**77,2 ± 0,2**	58,7 ± 0,3	**66,7 ± 0,2**

TABLE 3 – Comparaison des combinaisons de CamemBERT pré-entraîné sur tous les corpus, QuickMatching et Pyramid. Les résultats sont donnés sous la même forme que dans le tableau 2

Modèle	Test EMEA			Test MEDLINE		
	P	R	F1	P	R	F1
(Afzal *et al.*, 2016)	75,1	76,1	75.6	71,1	62,5	66,5
(van Mulligen *et al.*, 2016)	71,6	78,5	74.9	68,0	71,6	69,8
EDS-fine-tuned (Dura *et al.*, 2022)	-	-	72,9	-	-	59,7
(Chernyshevich & Stankevitch, 2015)	85,8	59,7	70.4	76,1	40,1	52,6

TABLE 4 – État de l'art sur le jeu de données QUAERO (métrique SeqEval pour Dura *et al.* (2022), BRAT-Eval ehealth pour les autres)

ment (jusqu'à 15 points sur MEDLINE pour l'union). Nous faisons l'hypothèse que l'amélioration de la couverture grâce à la combinaison des entités issues des trois approches en est la cause. En revanche, les deux méthodes diffèrent sur la précision : elle diminue fortement pour l'union, tandis qu'elle augmente pour le vote, de 5 et 14 points pour EMEA et MEDLINE respectivement. Cela résulte de la sélection des entités, présente dans le vote mais pas dans l'union.

Pour les experts et d'éventuelles applications concrètes dans le domaine de la santé, la difficulté réside dans le juste équilibre entre la précision et le rappel pour les cas d'usage précités.

Comparaison avec l'état de l'art Pour finir, nous comparons nos résultats avec les résultats de l'état de l'art, obtenus principalement lors des campagnes d'évaluation CLEF eHealth. Si nos méthodes peuvent rivaliser avec certains systèmes, comme celui de Chernyshevich & Stankevitch (2015) pour les deux corpus, il faut remarquer qu'une approche très fortement fondée sur des dictionnaires complétés par une traduction de termes en anglais, en l'occurrence (Afzal *et al.*, 2016), obtient de bien meilleurs résultats que QuickMatching, laissant à penser que la couverture de nos terminologies est insuffisante. Sur un autre plan, les performances de Chernyshevich & Stankevitch (2015), qui mettent en œuvre la fusion d'un grand nombre de modèles de type Champs Aléatoires Conditionnels (CRF), nous poussent à considérer des stratégies de fusion plus élaborées que celle que nous avons expérimentée. Finalement, nous pouvons constater que l'utilisation de documents cliniques pour continuer le pré-entraînement d'un modèle CamemBERT (Dura *et al.*, 2022) permet un gain significatif par rapport à nos méthodes.

4 Analyse des concepts extraits pour l'orthophonie

Le vote (cf. section 2.3) a ensuite été appliqué sur OrthoCorpus (2019), un corpus contenant des articles de la revue spécialisée *Rééducation Orthophonique*. Afin d'analyser plus facilement les concepts ainsi extraits (cf. section 4.2), nous proposons un nouveau type de qualification, les patrons syntaxico-sémantiques, que nous présentons dans la section 4.1, associés à une évaluation qualitative

de leurs ambiguïtés sur le corpus d'entraînement de QUAERO.

4.1 Patrons syntaxico-sémantiques des concepts médicaux

Nous proposons dans un premier temps une exploration de la structure linguistique de formation des termes médicaux, sous la forme de l'extraction de patrons syntaxico-sémantiques caractérisant ces concepts.

Plus précisément, les noms et adjectifs de l'UMLS sont d'abord extraits, associés à leurs types sémantiques. D'autres classes d'adjectifs sont également ajoutées, comme les adjectifs de quantification. Les termes complexes sont alors représentés comme des patrons syntaxico-sémantiques associant les catégories grammaticales et les types sémantiques de l'UMLS : par exemple, on identifie le patron *NOUN_diso ADJ_anat* indiquant un nom de pathologie suivi d'un adjectif anatomique. Ces patrons sont alors associés à des types sémantiques spécifiques en les projetant dans une annotation de référence. La figure 1 montre ainsi que certains patrons sont peu ambigus alors que d'autres, comme *NOUN_proc ADJ_anat*, se retrouvent en pratique dans des concepts de types différents.

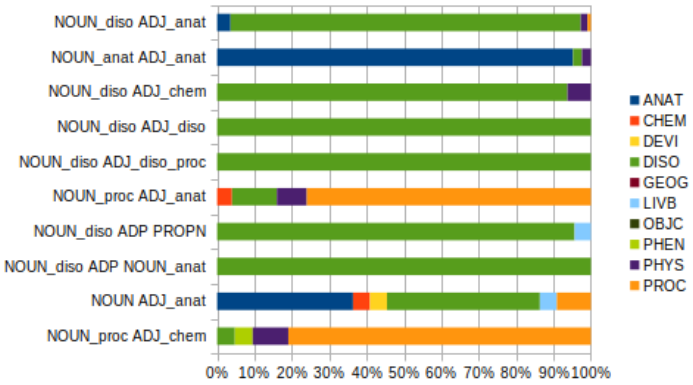


FIGURE 1 – Statistiques de projection de patrons syntaxico-sémantiques sur le corpus d'entraînement de QUAERO

En plus de fournir une analyse linguistique plus fine de la structuration syntaxico-sémantique des concepts médicaux, ces patrons sont également exploitables pour l'extraction automatique de concepts en santé. Ils offrent la possibilité de supprimer des annotations incohérentes avec les patrons, d'étendre des annotations trouvées selon les patrons ou de modifier les annotations selon les fréquences d'association des patrons avec les classes dans l'ensemble d'apprentissage. Dans nos expérimentations sur le corpus QUAERO, ces stratégies n'améliorent pas les résultats sur EMEA mais permettent de gagner jusqu'à deux points de f-score sur MEDLINE.

4.2 Analyse qualitative des candidats-termes classés

Un ensemble de plus de 11 000 candidats-termes a été collecté, dont plus de la moitié d'expressions polylexicales (Candito *et al.*, 2020), organisés selon dix des types sémantiques de l'UMLS. Une première analyse a montré que seuls certains de ces types ont une pertinence réelle du point de vue de la pratique de l'orthophonie (Anatomie, Pathologies, Physiologie, Procédures, Dispositifs, Êtres vivants).

De plus, on identifie bien certaines des spécificités du domaine : par exemple, on trouve *langue*, *cérébral*, *cordes vocales* et *système nerveux* parmi les termes d'Anatomie les plus fréquents, *trouble*

du langage et *trouble de la déglutition* parmi les Pathologies les plus citées, même si cette catégorie est très générale et mêle étiologie (*TC*, *lésion cérébrale*), symptômes (*modifications posturales*, *refus alimentaire*) et résultats (*échec scolaire*).

Par ailleurs, le problème de l'homonymie des termes est particulièrement présent, pour des raisons sémantiques mais également parce qu'aucune annotation syntaxique n'a été ajoutée. Ainsi le terme « marqueurs », identifié comme Nom est classé dans les produits chimiques, alors qu'il fait référence à des notions linguistiques (« des marqueurs du nom ») ou cliniques (« la présence de dysarthrie est un marqueur de mauvais pronostic »).

Cette analyse montre également qu'il reste beaucoup de bruit et d'imprécision dans l'extraction des termes (avec des termes manquants ou partiels) et qu'un travail manuel de sélection resterait nécessaire pour répondre aux besoins identifiés. En effet, les professionnels de santé (dont les orthophonistes) ont besoin d'explorer des documents non structurés de spécialité afin de réfléchir à leurs pratiques cliniques ainsi que de sélectionner et inclure des patients dans les études cliniques. Pour ce faire, la sélection fiable de termes pertinents est indispensable. La combinaison de méthodes automatiques et manuelles reste primordiale pour assurer une validité de la démarche.

5 Conclusions et perspectives

Nous avons présenté une approche hybride pour annoter des entités nommées dans le domaine médical. Cette approche combine un annotateur fondé sur des dictionnaires, un modèle de langue neuronal adapté au domaine avec des corpus de taille réduite et un modèle neuronal permettant de générer des entités imbriquées. Par ailleurs, pour ce qui est des modèles de langue neuronaux, nous avons montré qu'il n'est pas nécessaire d'avoir des corpus de grande taille pour observer une amélioration des résultats par rapport à un modèle dont le domaine n'a pas été adapté. Des études plus approfondies sur la similarité entre les corpus de test et les corpus utilisés pour l'adaptation permettront d'analyser plus finement ces résultats.

À plus long terme, nous continuerons à améliorer les modèles neuronaux par pré-entraînement sur des corpus non annotés ainsi qu'à enrichir les ressources de notre approche par dictionnaire. L'accès à des ressources de spécialité (Dictionnaire d'orthophonie) permettra également d'enrichir les tests sur ce domaine. Nous souhaitons également adapter le modèle CamemBERT aux entités imbriquées afin d'en améliorer la couverture. Enfin, le corpus DEFT 2020 (Cardon *et al.*, 2020) étant constitué de cas cliniques en français, nous souhaiterions l'utiliser pour tester les méthodes présentées afin d'évaluer leur potentiel d'adaptabilité à un type de corpus différent. Cela nous permettrait également de comparer les résultats ainsi obtenus à ceux de Copara *et al.* (2020), qui utilisent des méthodes similaires.

Remerciements

Ces travaux ont bénéficié d'un financement dans le cadre du programme e-Meuse Santé, porté par le Département de la Meuse et soutenu par les Départements de la Haute-Marne et de la Meurthe et Moselle, les GIP Objectif Meuse et Haute-Marne, la Région Grand Est, l'Agence Régionale de Santé Grand Est, et la Banque des Territoires au titre du programme France 2030. Ils ont été réalisés grâce au supercalculateur Factory-IA financé par le Conseil Régional d'Ile-de-France.

Références

AFZAL Z., AKHONDI S., HAAGEN H., VAN MULLIGEN E. M. & KORS J. (2016). Concept Recognition in French Biomedical Text Using Automatic Translation. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, volume 9822, p. 162–173. DOI : 10.1007/978-3-319-44564-9_13.

AKIBA T., SANO S., YANASE T., OHTA T. & KOYAMA M. (2019). Optuna : A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

ANSM (2021). Base de données publique des médicaments.

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, **5**, 135–146. DOI : 10.1162/tacl_a_00051.

BRIN-HENRY F. (2018). Pour une harmonisation de la terminologie orthophonique : contribution du projet OrthoCorpus (2015- 2017). In *Terminologica. TOTh 2018*.

CANDITO M., CONSTANT M., RAMISCH C., SAVARY A., GUILLAUME B., PARMENTIER Y. & CORDEIRO S. R. (2020). A French corpus annotated for multiword expressions and named entities. *Journal of Language Modelling*, **8**(2), 415–479. DOI : 10.15398/jlm.v8i2.265, HAL : hal-03016721.

CARDON R., GRABAR N., GROUIN C. & HAMON T. (2020). Présentation de la campagne d'évaluation DEFT 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques. In R. CARDON, N. GRABAR, C. GROUIN & T. HAMON, Éds., *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). Atelier DÉfi Fouille de Textes*, p. 1–13, Nancy, France : ATALA.

CHERNYSHEVICH M. & STANKEVITCH V. (2015). IHS-RD-BELARUS : Clinical Named Entities Identification in French Medical Texts. In *CLEF*.

COPARA J., KNAFOU J., NADERI N., MORO C., RUCH P. & TEODORO D. (2020). Contextualized French language models for biomedical named entity recognition. 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). Atelier DÉfi Fouille de Textes*, p. 36–48, Nancy, France : ATALA et AFCP.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT 2019*.

DURA B., JEAN C., TANNIER X., CALLIGER A., BEY R., NEURAZ A. & FLICOTEAUX R. (2022). Learning structures of the french clinical language : development and validation of word embedding models using 21 million clinical reports from electronic health records. *ArXiv*, **abs/2207.12940**.

EMBAREK M. & FERRET O. (2008). Learning patterns for building resources about semantic relations in the medical domain. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco : European Language Resources Association (ELRA).

GAO S., KOTEVSKA O., SOROKINE A. & CHRISTIAN J. (2021). A pre-training and self-training approach for biomedical named entity recognition. *PLOS ONE*, **16**.

GURURANGAN S., MARASOVIĆ A., SWAYAMDIPTA S., LO K., BELTAGY I., DOWNEY D. & SMITH N. A. (2020). Don't Stop Pretraining : Adapt Language Models to Domains and Tasks. In *ACL 2020*.

KINGMA D. P. & BA J. (2015). Adam : A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, USA.

L'HOMME M.-C. (2011). Y a-t-il une langue de spécialité ? points de vue pratique et théorique. *Langues et linguistique*, p. 26–33.

LINDBERG D. A., HUMPHREYS B. L. & MCCRAY A. T. (1993). The unified medical language system. *Methods of information in medicine*, **32**(4), 281–291.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty French language model. In *ACL 2020*.

NÉVÉOL A., COHEN K. B., GROUIN C., HAMON T., LAVERGNE T., KELLY L., GOEURIOT L., REY G., ROBERT A., TANNIER X. & ZWEIGENBAUM P. (2016). Clinical information extraction at the clef ehealth evaluation lab 2016. *CEUR workshop proceedings*, **1609**, 28—42.

NÉVÉOL A., GROUIN C., LEIXA J., ROSSET S. & ZWEIGENBAUM P. (2014). The QUAERO French medical corpus : A ressource for medical entity recognition and normalization. In *LREC Workshop BioTxtM2014*, p. 24–30.

OKAZAKI N. & TSUJII J. (2010). Simple and Efficient Algorithm for Approximate Dictionary Matching. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, p. 851–859, Beijing, China : Coling 2010 Organizing Committee.

ORTHOCORPUS (2019). ATILF. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

SOLDAINI L. & GOHARIAN N. (2016). QuickUMLS : a fast, unsupervised approach for medical concept extraction. In *SIGIR MedIR workshop*, p. 1–4.

VAN MULLIGEN E. M., AFZAL Z., AKHONDI S., VO-HAI D. & KORS J. A. (2016). Erasmus MC at CLEF eHealth 2016 : Concept recognition and coding in French texts. In *CEUR Workshop Proceedings*, p. 171–178 : CLEF.

VERSPOOR K., JIMENO YEPES A., CAVEDON L., MCINTOSH T., HERTEN-CRABB A., THOMAS Z. & PLAZZER J.-P. (2013). Annotating the biomedical literature for the human variome. *Database*, **2013**.

WANG J., SHOU L., CHEN K. & CHEN G. (2020). Pyramid : A layered model for nested named entity recognition. In *ACL 2020* : Association for Computational Linguistics.

WEI X., WANG S., ZHANG D., BHATIA P. & ARNOLD A. (2021). Knowledge Enhanced Pretrained Language Models : A Comprehensive Survey. *arXiv :2110.08455 [cs]*.

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M., DAVISON J., SHLEIFER S., VON PLATEN P., MA C., JERNITE Y., PLU J., XU C., SCAO T. L., GUGGER S., DRAME M., LHOEST Q. & RUSH A. M. (2020). Transformers : State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 38–45, Online : Association for Computational Linguistics.

YANG J., XIAO G., SHEN Y., JIANG W., HU X., ZHANG Y. & PENG J. (2022). A Survey of Knowledge Enhanced Pre-trained Models. *arXiv :2110.00269 [cs]*.

YIN D., DONG L., CHENG H., LIU X., CHANG K.-W., WEI F. & GAO J. (2022). A Survey of Knowledge-Intensive NLP with Pre-Trained Language Models. *arXiv :2202.08772 [cs]*.

Facilitating NLP specialists' access to language archive materials: an update

Benjamin Galliot[1] Guillaume Wisniewski[2] Séverine Guillaume[1]
Guillaume Jacques[3] Alexis Michaud[1]
(1) Langues et Civilisations à Tradition Orale (LACITO), CNRS - Sorbonne Nouvelle - INALCO
(2) Laboratoire de Linguistique Formelle (LLF), CNRS - Université de Paris
(3) Centre de Recherches Linguistiques sur l'Asie Orientale (CRLAO), 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, rgyalrongskad@gmail.com,
alexis.michaud@cnrs.fr

Abstract

We present a software tool to assemble a great range of diverse datasets from the Pangloss collection (a multimedia open archive of under-documented languages). The tool ensures the reproducibility of experiments conducted on these data. As an example, two transcribed audio corpora of Chinese minority languages (Japhug and Na) are proposed, under a Creative Commons license, as reference corpora for experiments in Natural Language Processing, and as examples of a pipeline that can be generalized to other corpora from open archives. An overarching goal of making language archive data available in an easily accessible and usable form is to facilitate the development and deployment of state-of-the-art natural language processing tools for the full range of human languages. This presentation, which follows a previous paper on the same topic, reports on new developments including feedback on a deposit at Hugging Face Datasets.

Résumé

Faciliter l'accès des praticiens du traitement automatique des langues à des jeux de données de langues rares : un deuxième point d'étape

Nous présentons un outil logiciel qui permet d'assembler divers jeux de données de la collection Pangloss (archive ouverte multimédia de langues rares) en assurant la reproductibilité des expériences menées sur ces données. À titre d'exemple, deux corpus audio transcrits de langues minoritaires de Chine (japhug et na) sont proposés, sous une licence Creative Commons, comme corpus de référence pour des expériences en traitement automatique des langues, et comme exemples d'une chaîne de traitement généralisable à d'autres corpus d'archives ouvertes. L'enjeu global d'une mise à disposition de données de langues rares sous une forme aisément accessible et utilisable est de faciliter le développement et le déploiement d'outils de pointe en traitement automatique des langues naturelles pour tout l'éventail des langues humaines. Cet exposé, qui fait suite à une précédente communication sur le même thème, fait état de nouveautés dont un retour d'expérience concernant un dépôt auprès de Hugging Face.

Keywords: benchmark datasets, computational language documentation, endangered languages.

Mots-clés : corpus de référence, documentation computationnelle des langues, langues rares.

Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain *(LIFT)* et Traitement Automatique des Langues *(TAL)*,
Marseille, France, 14 au 15 novembre 2022 , pages 109–118,

1 Introduction

The deployment of automatic speech processing tools clearly has important implications for language documentation, at a time when the decline in linguistic diversity is accelerating (Kik et al., 2021), in parallel with the decline in biodiversity. Conversely, less-documented languages present a whole range of challenges to computational research, the interest of which is increasingly clearly perceived (Anastasopoulos et al., 2020). In this context, providing easily accessible, clearly versioned and user-friendly corpora of less-studied languages appears as a central necessity. This presentation, which follows a previous paper on the same theme (Galliot et al., 2021), presents our contribution to this endeavour. Following the example of the publication of the Mbochi (Bantu) corpus (Godard et al., 2018), we have deposited in Zenodo and in Hugging Face Datasets audio corpora (with transcriptions) of Japhug and Na, two minority languages of China.

These corpora have been used in automatic speech recognition work (Adams et al., 2018; Adams et al., 2021; Guillaume, Wisniewski, Macaire, et al., 2022; Macaire, 2021) and in interdisciplinary reflections associating NLP specialists and linguists (Guillaume, Wisniewski, Galliot, et al., 2022; Michaud et al., 2018 ; Michaud et al., 2020). The corpora are available online in an open archive, the Pangloss Collection[1] (Michaud et al., 2016), an open archive of (mostly) endangered languages, itself hosted by the Cocoon data repository[2], ensuring availability in open access.

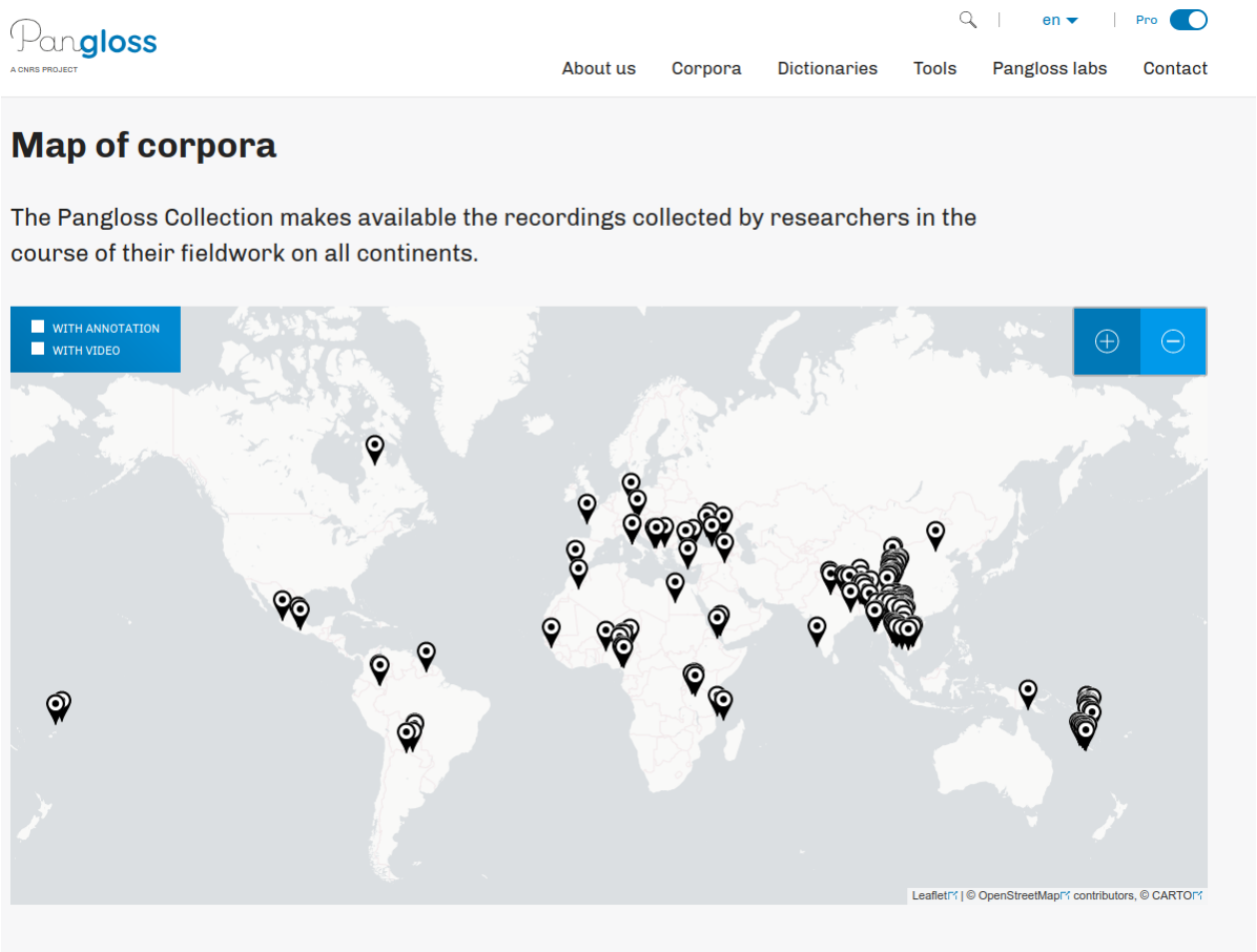


Figure 1: Pangloss – map of languages

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

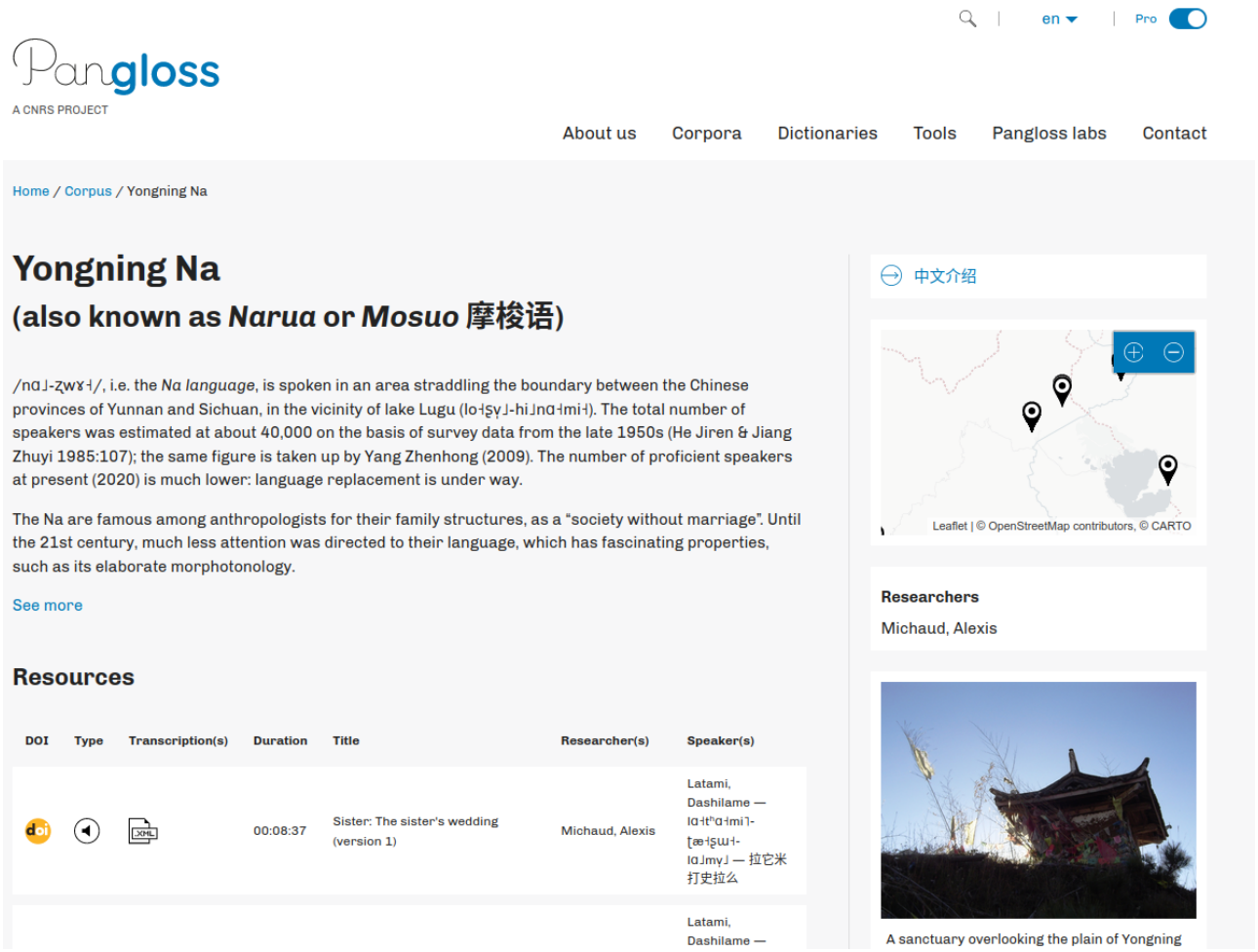


Figure 2: Pangloss – page of a corpus: introduction and list of available multimedia resources

The transcriptions of these corpora are enriched and revised over the years, however, and new documents are added, so that referring to the open archive itself is not a sufficiently precise reference to achieve actual reproducibility of the experiments conducted on these data. Even though the primary data (the audio and video files) do not change, the transcriptions constitute "lukewarm data", which get improved and modified slowly over time.

We have therefore made a deposit of a given state of these two corpora, making them accessible in a few clicks, in a stabilized form, in Zenodo (§3) and among the Hugging Face Datasets (§4). But the most important advance, in our view, is the creation of a script, OutilsPangloss (§2), to select among the corpora of the Pangloss Collection while maintaining a guarantee of full reproducibility (thanks to the versioning system of documents deposited in the archive).

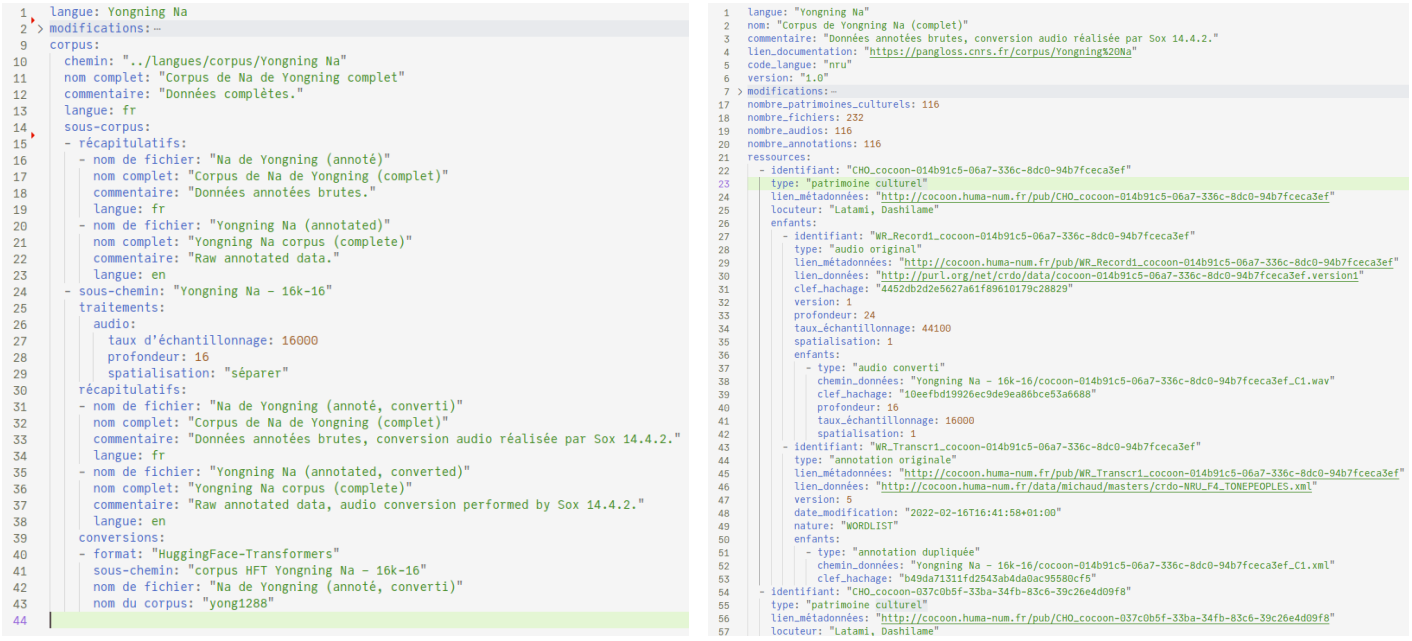
2 A tool to build new datasets

The tool developed during the preparation of the corpora deposited in Zenodo and Hugging Face, entitled OutilsPangloss[3], allows for putting together an open-ended range of new datasets. It consists of a toolbox in Julia language, which, among other functionalities, creates (sub)corpora of less-documented language data from the Pangloss Collection. The user fills in a YAML file (of which various examples are provided, see Figure 3a), specifying the names of the desired languages (as they

[3] https://gitlab.com/lacito/outilspangloss

appears in the Pangloss Collection). It can also provide a list of regular expressions for modifications in case processing of annotations is to be carried out (such as deletions or rearrangements of text blocks). It is possible to filter by speaker. Processing on the audio can also be set, to choose the sampling rate and bit-depth, and to separate the different tracks of multi-channel files into mono files (demultiplexing).

After harvesting (in Sparql), metadata checks (hashing, versions, etc.), data checks (mistake detection, etc.), and downloads, a general summary file (see Figure 3b) is generated in the target folder, alongside data and metadata folders (see Figure 4). The information contained in this summary file is sufficient to reproduce exactly, at any time, an experiment conducted with the dataset it describes.



```yaml
langue: Yongning Na
modifications:…
corpus:
  chemin: "../langues/corpus/Yongning Na"
  nom complet: "Corpus de Na de Yongning complet"
  commentaire: "Données complètes."
  langue: fr
  sous-corpus:
  - récapitulatifs:
    - nom de fichier: "Na de Yongning (annoté)"
      nom complet: "Corpus de Na de Yongning (complet)"
      commentaire: "Données annotées brutes."
      langue: fr
    - nom de fichier: "Yongning Na (annotated)"
      nom complet: "Yongning Na corpus (complete)"
      commentaire: "Raw annotated data."
      langue: en
  - sous-chemin: "Yongning Na – 16k-16"
    traitements:
      audio:
        taux d'échantillonnage: 16000
        profondeur: 16
        spatialisation: "séparer"
    récapitulatifs:
    - nom de fichier: "Na de Yongning (annoté, converti)"
      nom complet: "Corpus de Na de Yongning (complet)"
      commentaire: "Données annotées brutes, conversion audio réalisée par Sox 14.4.2."
      langue: fr
    - nom de fichier: "Yongning Na (annotated, converted)"
      nom complet: "Yongning Na corpus (complete)"
      commentaire: "Raw annotated data, audio conversion performed by Sox 14.4.2."
      langue: en
    conversions:
    - format: "HuggingFace-Transformers"
      sous-chemin: "corpus HFT Yongning Na – 16k-16"
      nom de fichier: "Na de Yongning (annoté, converti)"
      nom du corpus: "yong1288"
```

(a) corpus configuration

```yaml
langue: "Yongning Na"
nom: "Corpus de Yongning Na (complet)"
commentaire: "Données annotées brutes, conversion audio réalisée par Sox 14.4.2."
lien_documentation: "https://pangloss.cnrs.fr/corpus/Yongning%20Na"
code_langue: "nru"
version: "1.0"
modifications:…
nombre_patrimoines_culturels: 116
nombre_fichiers: 232
nombre_audios: 116
nombre_annotations: 116
ressources:
  - identifiant: "CHO_cocoon-014b91c5-06a7-336c-8dc0-94b7fceca3ef"
    type: "patrimoine culturel"
    lien_métadonnées: "http://cocoon.huma-num.fr/pub/CHO_cocoon-014b91c5-06a7-336c-8dc0-94b7fceca3ef"
    locuteur: "Latami, Dashilame"
    enfants:
      - identifiant: "WR_Record1_cocoon-014b91c5-06a7-336c-8dc0-94b7fceca3ef"
        type: "audio original"
        lien_métadonnées: "http://cocoon.huma-num.fr/pub/WR_Record1_cocoon-014b91c5-06a7-336c-8dc0-94b7fceca3ef"
        lien_données: "http://purl.org/net/crdo/data/cocoon-014b91c5-06a7-336c-8dc0-94b7fceca3ef.version1"
        clef_hachage: "4452db2d2e5627a61f89610179c28829"
        version: 1
        profondeur: 24
        taux_échantillonnage: 44100
        spatialisation: 1
        enfants:
          - type: "audio converti"
            chemin_données: "Yongning Na – 16k-16/cocoon-014b91c5-06a7-336c-8dc0-94b7fceca3ef_C1.wav"
            clef_hachage: "10eefbd19926ec9de9ea86bce53a6688"
            profondeur: 16
            taux_échantillonnage: 16000
            spatialisation: 1
      - identifiant: "WR_Transcr1_cocoon-014b91c5-06a7-336c-8dc0-94b7fceca3ef"
        type: "annotation originale"
        lien_métadonnées: "http://cocoon.huma-num.fr/pub/WR_Transcr1_cocoon-014b91c5-06a7-336c-8dc0-94b7fceca3ef"
        lien_données: "http://cocoon.huma-num.fr/data/michaud/masters/crdo-NRU_F4_TONEPEOPLES.xml"
        version: 6
        date_modification: "2022-02-16T16:41:58+01:00"
        nature: "WORDLIST"
        enfants:
          - type: "annotation dupliquée"
            chemin_données: "Yongning Na – 16k-16/cocoon-014b91c5-06a7-336c-8dc0-94b7fceca3ef_C1.xml"
            clef_hachage: "b49da71311fd2543ab4da0ac95580cf5"
  - identifiant: "CHO_cocoon-037c0b5f-33ba-34fb-83c6-39c26e4d09f8"
    type: "patrimoine culturel"
    lien_métadonnées: "http://cocoon.huma-num.fr/pub/CHO_cocoon-037c0b5f-33ba-34fb-83c6-39c26e4d09f8"
    locuteur: "Latami, Dashilame"
```

(b) Versioned corpus (list of resources, metadata...)

Figure 3: OutilsPangloss : YML files

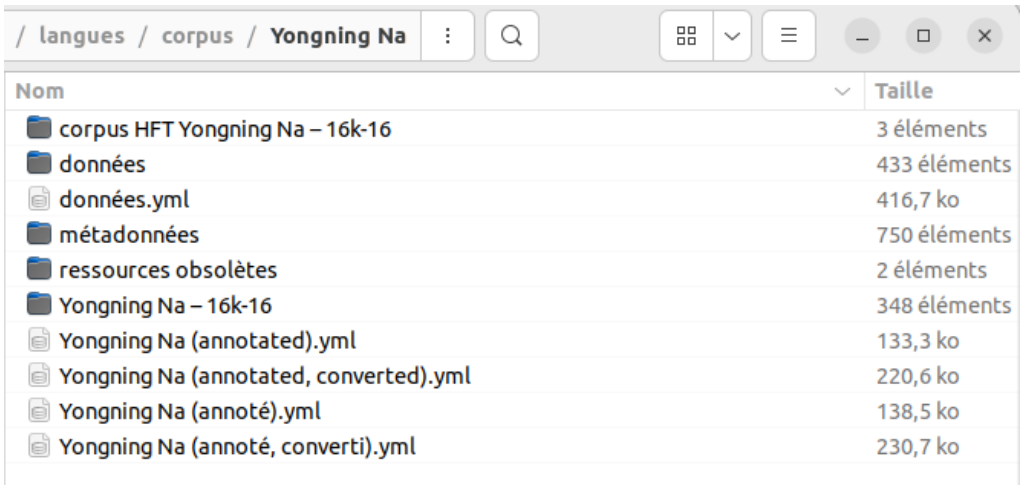
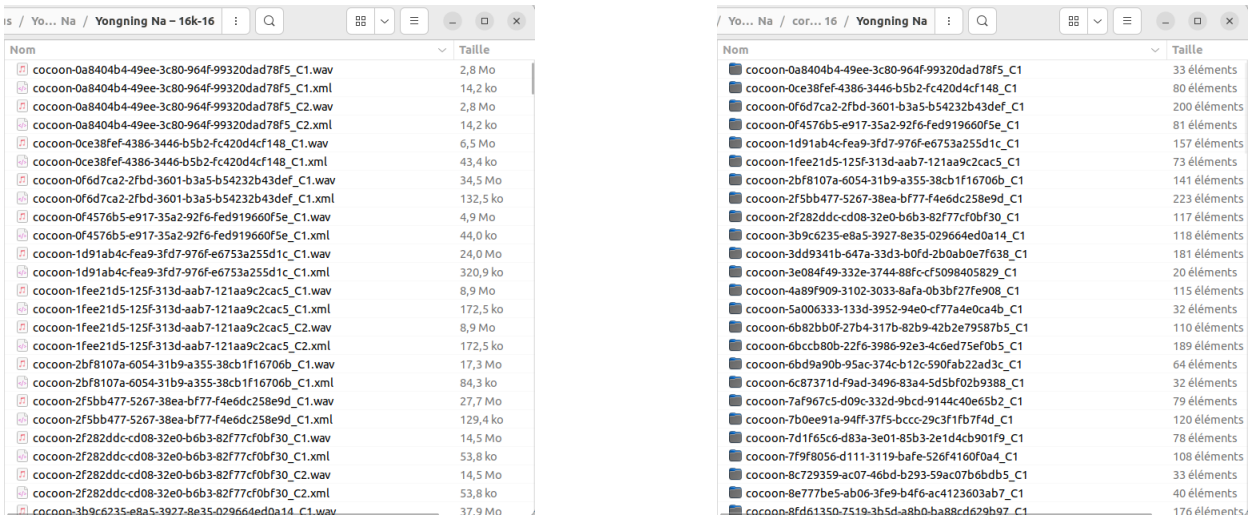


Figure 4: OutilsPangloss: local target folder of a corpus



(a) processed resources

(b) processed resources for the Hugging Face format

Figure 5: OutilsPangloss : données

3 The Zenodo deposits: access links and explanations about technical choices

Links to the repository The Na and Japhug corpora were uploaded to Zenodo, where they constitute deposits 5336698 (Na) and 5521112 (Japhug), respectively. An entire corpus is identified by a DOI (*digital object identifier*): `10.5281/zenodo.5521112` for the Japhug corpus, and `10.5281/zenodo.5336698` for the Na corpus.

The same type of identifier has been deployed for the Pangloss Collection, the open archive where the corpora are deposited, but with a completely different granularity: a DOI for each document (Vasile et al., 2020), a choice which is suitable for linguists who wish to reference data at a highly specific level (a text and, within a text, a specific utterance) but does not give a handle on an entire corpus.

Audio files Audio files were downgraded to 16-bit, 16 kHz, mono (see Figure 5a). The logic behind this choice of parameters is that NLP experiments have requirements that differ from those of long-term archiving in the Pangloss Collection. The size of the two datasets deposited in Zenodo is compatible (as of today) with experiments carried out on a laptop computer: 1.8 GB for Na, 9.2 GB for Japhug.

Annotations The annotations are in the original format: XML structured according to a simple hierarchy (a text is composed of sentences, themselves composed of words, themselves composed of morphemes). An example is provided in Figure 6.

Some basic preprocessing has been carried out, so that users do not have to learn about the various conventions chosen by the depositors, which vary across corpora. In particular, when transcribing texts, it is not infrequent for language consultants and language workers (typically linguists) to make decisions that create an edit distance between the transcription and what is actually said in the recording. A convention used in the Na corpus is that added passages are placed within square brackets. Another convention is that passages that the language consultants wish to see removed from the "smoothed" transcription are placed between angle brackets. At preprocessing, the passages within square brackets were deleted, and the angle brackets were removed, thereby ensuring the closest match between audio and transcriptions.

(a) data (b) metadata

Figure 6: XML files of annotation resources

4 The Pangloss corpus in Hugging Face Datasets

Hugging Face (HF) Datasets are currently a key hub for NLP researchers looking for easily usable corpora. The formatting of Pangloss corpora in HF format therefore seemed useful from the perspective of bringing corpora from less-studied languages to the attention of NLP researchers.

Preparing the data for Hugging Face Datasets format Each annotation file is divided into as many files as there are first-level units (sentences for texts, words for vocabulary lists). These files are named after these units, and are placed in folders named after the original resource (see Figure 5b). These data, structured in a tabular manner, contain some relevant metadata, such as the nature of the resource (word list or text) or the speaker. The data are then randomly partitioned. By default, the partitioning is carried out at the first level below the entire resource: hence sentences for texts, and words in the case of word lists. But it is possible to choose the level of the entire file instead. Three sets are thereby created: training, validation and test, thus three CSV files (see figure 7). Finally, all these data (audio files of the sentences and the three CSV files) are uploaded on a server.

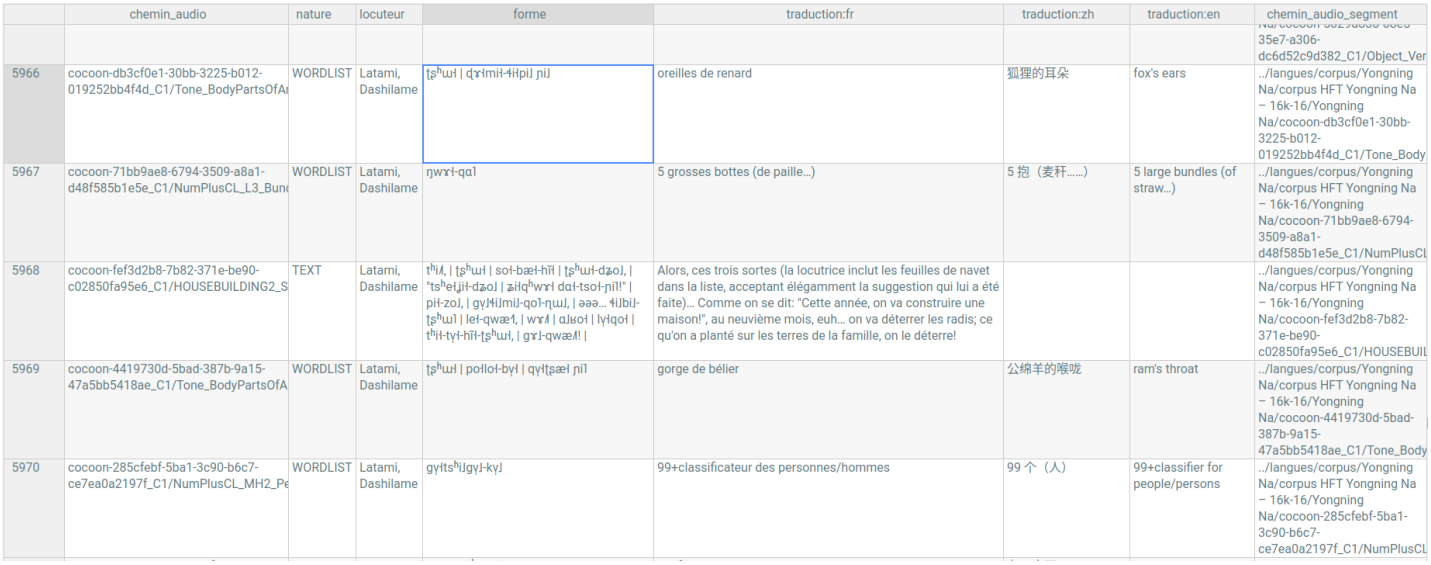
Preparing the publicly available dataset A second part of the work consists in formally preparing the new datasets[4]. To achieve this, a precise description of each corpus is required. Particular attention is paid to the encoding of the language concerned[5], a delicate point in the case of languages that are not standardized (or only slightly standardized), which constitute the core business of the Pangloss Collection. A Python script[6] automates the creation of the dataset from the previously uploaded and accessible corpus archives. This script, of varying complexity depending on the dataset, is used to specify where the formatted archives are stored, the data types of each column, and their corresponding names. It is then possible to view the corpus data directly online and to listen to the audio segments (see Figure 8).

The choice made consists in creating one global dataset named Pangloss, of which the various

[4]Available here: https://huggingface.co/datasets/Lacito/pangloss

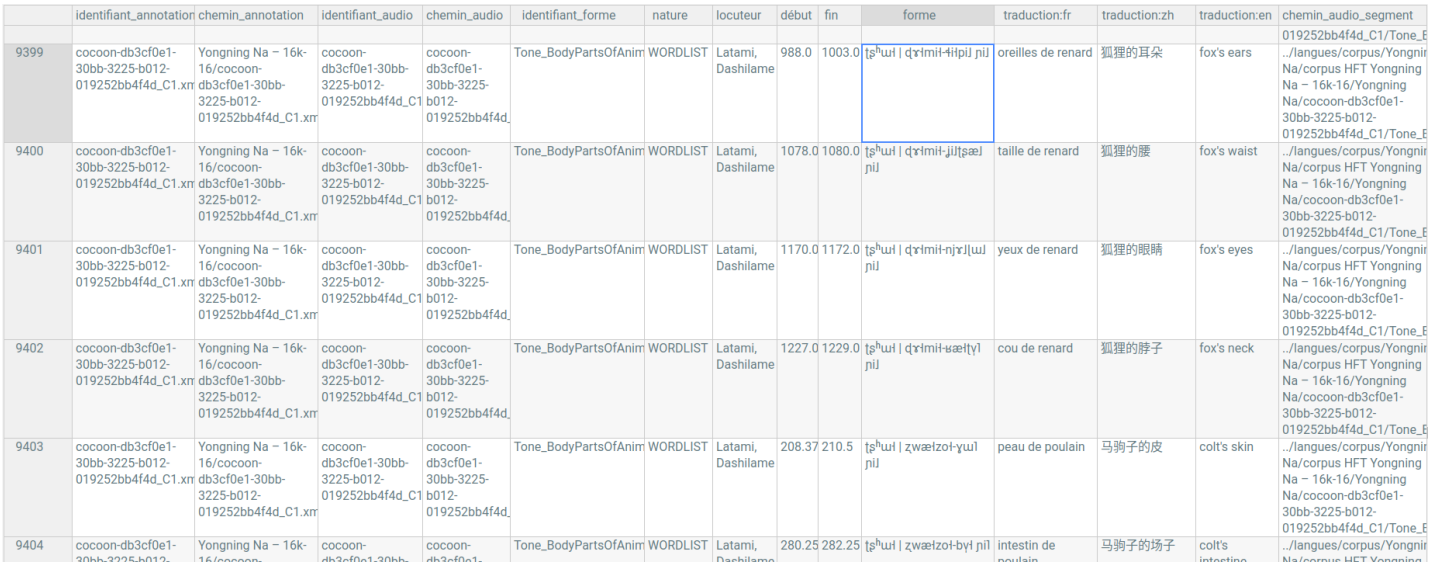
[5]https://github.com/huggingface/datasets/issues/4881

[6]https://huggingface.co/datasets/Lacito/pangloss/blob/main/pangloss.py



	chemin_audio	nature	locuteur	forme	traduction:fr	traduction:zh	traduction:en	chemin_audio_segment
								35e7-a306-dc6d52c9d382_C1/Object_Ver
5966	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1/Tone_BodyPartsOfA	WORDLIST	Latami, Dashilame	tʂʰɯɨ | ɖɤɪmiɬ-ɬiɬpiJ ɲiJ	oreilles de renard	狐狸的耳朵	fox's ears	../langues/corpus/Yongning Na/corpus HFT Yongning Na – 16k-16/Yongning Na/cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1/Tone_Body
5967	cocoon-71bb9ae8-6794-3509-a8a1-d48f585b1e5e_C1/NumPlusCL_L3_Bun	WORDLIST	Latami, Dashilame	ŋwɤɪ-qɑ1	5 grosses bottes (de paille...)	5 抱（麦秆……）	5 large bundles (of straw...)	../langues/corpus/Yongning Na/corpus HFT Yongning Na – 16k-16/Yongning Na/cocoon-71bb9ae8-6794-3509-a8a1-d48f585b1e5e_C1/NumPlusCL
5968	cocoon-fef3d2b8-7b82-371e-be90-c02850fa95e6_C1/HOUSEBUILDING2_S	TEXT	Latami, Dashilame	tʰiʎ, | tʂʰɯɨ | soɬ-bæɬ-hĩɪ̃ | tʂʰɯɨ-dʑoJ, | "tsʰeɟiɬ-dʑoJ | ʑiɬqʰwɤɪ dɑɬ-tsoɬ-ɲĩ1!" | piɬ-zoJ, | gʏɬiʟmiJ-qoɬ-ɳɯJ, | əəə... ɬiʟbiJ-tʂʰɯJ | leɬ-qwæ1, | wɤɻ | ɑɻoɬ | lʏɬqoɬ | tʰiɬ-tʏɬ-hĩɪ̃-tʂʰɯɨ, | gɤɻ-qwæɻ! |	Alors, ces trois sortes (la locutrice inclut les feuilles de navet dans la liste, acceptant élégamment la suggestion qui lui a été faite)... Comme on se dit: "Cette année, on va construire une maison!", au neuvième mois, euh... on va déterrer les radis; ce qu'on a planté sur les terres de la famille, on le déterre!			../langues/corpus/Yongning Na/corpus HFT Yongning Na – 16k-16/Yongning Na/cocoon-fef3d2b8-7b82-371e-be90-c02850fa95e6_C1/HOUSEBUIL
5969	cocoon-4419730d-5bad-387b-9a15-47a5bb5418ae_C1/Tone_BodyPartsOfA	WORDLIST	Latami, Dashilame	tʂʰɯɨ | poɬloɬ-bʏɬ | qʏɬʈʂæɬ ɲiJ	gorge de bélier	公绵羊的喉咙	ram's throat	../langues/corpus/Yongning Na/corpus HFT Yongning Na – 16k-16/Yongning Na/cocoon-4419730d-5bad-387b-9a15-47a5bb5418ae_C1/Tone_Body
5970	cocoon-285cfebf-5ba1-3c90-b6c7-ce7ea0a2197f_C1/NumPlusCL_MH2_Pe	WORDLIST	Latami, Dashilame	gʏtsʰiJgʏJ-kʏJ	99+classificateur des personnes/hommes	99 个（人）	99+classifier for people/persons	../langues/corpus/Yongning Na/corpus HFT Yongning Na – 16k-16/Yongning Na/cocoon-285cfebf-5ba1-3c90-b6c7-ce7ea0a2197f_C1/NumPlusCL

(a) data for the Hugging Face format



	identifiant_annotation	chemin_annotation	identifiant_audio	chemin_audio	identifiant_forme	nature	locuteur	début	fin	forme	traduction:fr	traduction:zh	traduction:en	chemin_audio_segment
														019252bb4f4d_C1/Tone_B
9399	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1.xm	Yongning Na – 16k-16/cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1.xm	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_	Tone_BodyPartsOfAnim	WORDLIST	Latami, Dashilame	988.0	1003.0	tʂʰɯɨ | ɖɤɪmiɬ-ɬiɬpiJ ɲiJ	oreilles de renard	狐狸的耳朵	fox's ears	../langues/corpus/Yongnin Na/corpus HFT Yongning Na – 16k-16/Yongning Na/cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1/Tone_B
9400	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1.xm	Yongning Na – 16k-16/cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1.xm	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_	Tone_BodyPartsOfAnim	WORDLIST	Latami, Dashilame	1078.0	1080.0	tʂʰɯɨ | ɖɤɪmiɬ-jiʟtʂæɻ ɲiJ	taille de renard	狐狸的腰	fox's waist	../langues/corpus/Yongnin Na/corpus HFT Yongning Na – 16k-16/Yongning Na/cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1/Tone_B
9401	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1.xm	Yongning Na – 16k-16/cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1.xm	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_	Tone_BodyPartsOfAnim	WORDLIST	Latami, Dashilame	1170.0	1172.0	tʂʰɯɨ | ɖɤɪmiɬ-njɤɻ[ɯɻ ɲiJ	yeux de renard	狐狸的眼睛	fox's eyes	../langues/corpus/Yongnin Na/corpus HFT Yongning Na – 16k-16/Yongning Na/cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1/Tone_B
9402	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1.xm	Yongning Na – 16k-16/cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1.xm	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_	Tone_BodyPartsOfAnim	WORDLIST	Latami, Dashilame	1227.0	1229.0	tʂʰɯɨ | ɖɤɪmiɬ-ʁæɬʈʏl ɲiJ	cou de renard	狐狸的脖子	fox's neck	../langues/corpus/Yongnin Na/corpus HFT Yongning Na – 16k-16/Yongning Na/cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1/Tone_B
9403	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1.xm	Yongning Na – 16k-16/cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1.xm	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_	Tone_BodyPartsOfAnim	WORDLIST	Latami, Dashilame	208.37	210.5	tʂʰɯɨ | ʐwæɬzoɬ-ɣɯɻ ɲiJ	peau de poulain	马驹子的皮	colt's skin	../langues/corpus/Yongnin Na/corpus HFT Yongning Na – 16k-16/Yongning Na/cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1/Tone_B
9404	cocoon-db3cf0e1-30bb-3225-b012-	Yongning Na – 16k-16/cocoon-	cocoon-	cocoon-	Tone_BodyPartsOfAnim	WORDLIST	Latami,	280.25	282.25	tʂʰɯɨ | ʐwæɬzoɬ-bʏɬ ɲiJ	intestin de poulain	马驹子的场子	colt's intestine	../langues/corpus/Yongnin Na/corpus HFT Yongning

(b) data in Hugging Face format, after random partitioning

Figure 7: OutilsPangloss: CSV files

languages (Japhug[7], and Yongning Na[8]) constitute sub-sets.

It is then possible to load the desired data set and get experiments started in just two lines of Python code.[9]

[7] https://huggingface.co/datasets/Lacito/pangloss/viewer/japh1234/train

[8] https://huggingface.co/datasets/Lacito/pangloss/viewer/yong1288/train

[9] `import datasets` then
`datasets.load_dataset("pangloss", "japh1234")` for Japhug, and
`datasets.load_dataset("pangloss", "yong1288")` for Yongning Na.

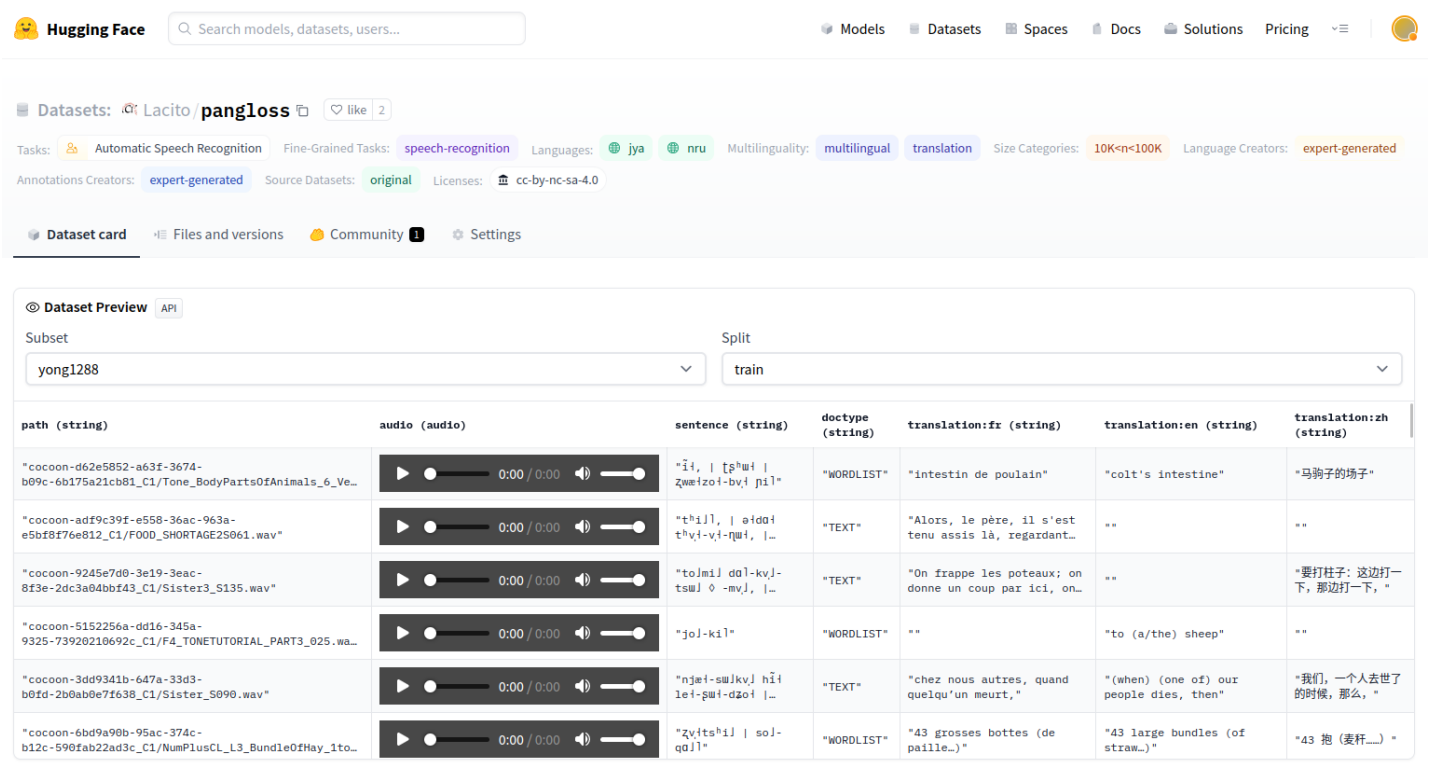


Figure 8: Page of the Pangloss corpus on Hugging Face Datasets

5 Perspectives: what is at stake in providing a full description of datasets without creating "hard copies"

The ongoing transition to Open Science carries a fundamental requirement for reproducibility of experiments, and the field of speech sciences and Automatic Language Processing is no exception (Garellek et al., 2020). Our hope is that practices will gradually evolve towards a description of datasets via metadata pointing to one master file hosted in an archive that guarantees both long-term conservation and 24/7 online availability. Describing the datasets in this way only takes a few kilobytes (Kb), whereas hosting each dataset as a "hard copy" would result in multiple and highly redundant deposits (in Zenodo or elsewhere) each of which amounts to gigabytes (GB).

Acknowledgments

Many thanks to the language consultants and friends for Japhug (in particular Tshendzin) and Na (in particular Mrs. Latami Dashilame and her son Latami Dashi). The present work is a contribution to the project "Computational language documentation by 2025" (ANR-19-CE38-0015-04) and to the Labex "Empirical foundations of linguistics" (ANR-10-LABX-0083). We thank the Institute for Linguistic Heritage and Diversity (ILARA) at *École pratique des hautes études*, the University of Queensland and the Australian Research Council Centre of Excellence for the Dynamics of Language for financial support to software development for language documentation.

References

ADAMS, O., COHN, T., NEUBIG, G., CRUZ, H., BIRD, S., & MICHAUD, A. (2018). Evaluating phonemic transcription of low-resource tonal languages for language documentation. *Proceedings of LREC 2018 (Language Resources and Evaluation Conference)*, 3356–3365. https://halshs.archives-ouvertes.fr/halshs-01709648

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., & HILL, N. (2021). User-friendly automatic transcription of low-resource languages: Plugging ESPnet into Elpis. *Proceedings of ComputEL-4: Fourth Workshop on the Use of Computational Methods in the Study of Endangered Languages*. https://halshs.archives-ouvertes.fr/halshs-03030529

ANASTASOPOULOS, A., COX, C., NEUBIG, G., & CRUZ, H. (2020). Endangered languages meet Modern NLP. *Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts*, 39–45. https://doi.org/10.18653/v1/2020.coling-tutorials.7

GALLIOT, B., WISNIEWSKI, G., GUILLAUME, S., MICHAUD, A., ROSSATO, S., NGUYÊN, M.-C., & FILY, M. (2021). Deux corpus audio transcrits de langues rares (japhug et na) normalisés en vue d'expériences en traitement du signal. *Journées scientifiques du Groupement de recherche "Linguistique informatique, formelle et de terrain" (GDR LIFT)*. https://halshs.archives-ouvertes.fr/halshs-03475436

GARELLEK, M., GORDON, M., KIRBY, J., LEE, W.-S., MICHAUD, A., MOOSHAMMER, C., NIEBUHR, O., RECASENS, D., ROETTGER, T. B., SIMPSON, A., et al. (2020). Toward open data policies in phonetics: What we can gain and how we can avoid pitfalls. *Journal of Speech Science*, *9*(1), 3–16.

GODARD, P., ADDA, G., ADDA-DECKER, M., BENJUMEA, J., BESACIER, L., COOPER-LEAVITT, J., KOUARATA, G. N., LAMEL, L., MAYNARD, H., & MUELLER, M. (2018). A very low resource language speech corpus for computational language documentation experiments. *Proceedings of LREC 2018 (Language Resources and Evaluation Conference)*, 3366–3370.

GUILLAUME, S., WISNIEWSKI, G., GALLIOT, B., NGUYỄN, M.-C., FILY, M., JACQUES, G., & MICHAUD, A. (2022). Plugging a neural phoneme recognizer into a simple language model: a workflow for low-resource settings. *Interspeech 2022 - 23rd Annual Conference of the International Speech Communication Association*. https://doi.org/10.5281/zenodo.5521111

GUILLAUME, S., WISNIEWSKI, G., MACAIRE, C., JACQUES, G., MICHAUD, A., GALLIOT, B., COAVOUX, M., ROSSATO, S., NGUYỄN, M.-C., & FILY, M. (2022). Fine-tuning pre-trained models for Automatic Speech Recognition: experiments on a fieldwork corpus of Japhug (Trans-Himalayan family). *ComputEL-5 5th Workshop on Computational Methods for Endangered Languages (ComputEL-5)*. https://halshs.archives-ouvertes.fr/halshs-03647315

KIK, A., ADAMEC, M., AIKHENVALD, A. Y., BAJZEKOVA, J., BARO, N., BOWERN, C., COLWELL, R. K., DROZD, P., DUDA, P., IBALIM, S., JORGE, L. R., MOGINA, J., RULI, B., SAM, K., SARVASY, H., SAULEI, S., WEIBLEN, G. D., ZRZAVY, J., & NOVOTNY, V. (2021). Language and ethnobiological skills decline precipitously in papua new guinea, the world's most linguistically diverse nation. *Proceedings of the National Academy of Sciences*, *118*(22). https://doi.org/10.1073/pnas.2100096118

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., ADAMS, O., COHN, T., NEUBIG, G., & 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., ADAMS, O., COX, C., GUILLAUME, S., WISNIEWSKI, G., & 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). https://halshs.archives-ouvertes.fr/halshs-02881731/

MICHAUD, A., GUILLAUME, S., JACQUES, G., MAC, D.-K., JACOBSON, M., PHAM, T.-H., & DEO, M. (2016). Contribuer au progrès solidaire des recherches et de la documentation : la Collection Pangloss et la Collection AuCo. *Journées d'Etude de la Parole 2016*, *1*, 155-163. https://halshs.archives-ouvertes.fr/halshs-01341631

VASILE, A., GUILLAUME, S., AOUINI, M., & 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. https://halshs.archives-ouvertes.fr/halshs-02870206

Génération de questions et paradigme question/réponse pour l'exploration des collections de sciences humaines numériques

Frédéric Béchet[1] Elie Antoine[1] Jeremy Auguste[1,3] Géraldine Damnati[2]
(1) Aix-Marseille Université, CNRS, LIS {first.last}@lis-lab.fr
(2) Orange Innovation, DATA&AI, Lannion {first.last}@orange.com
(3) IrAsia - Institut de recherches Asiatiques

RÉSUMÉ

Cet article présente notre méthode de génération de questions, proposant d'exploiter l'analyse sémantique de textes pour sélectionner des réponses plausibles et enrichir le processus de génération par des traits sémantiques génériques. Ces questions générées sont ensuite utilisées à plusieurs fins, d'une part comme une méthode d'adaptation de modèles de compréhension de documents, et de l'autre comme liens explicables entre documents dans le cadre d'une collection d'archives numérisées pour les études en sciences sociales. Un autre intérêt de cette étude est l'évaluation des méthodes de génération de questions et de compréhension de documents sur un nouveau type de documents, pour aller au-delà des évaluations de référence traditionnelles.

ABSTRACT

Question Generation and Answering for exploring Digital Humanities collections

This paper presents our method for generating questions, proposing to exploit semantic analysis of texts to select plausible answers and to enrich the generation process with generic semantic features. These generated questions are then used for several purposes, on the one hand as a method for adapting models of document understanding, and on the other hand as explainable links between documents in the context of a collection of digitized archives for social science studies. Another interest of this study is the evaluation of question generation and document comprehension methods on a new type of documents, to go beyond traditional reference evaluations.

MOTS-CLÉS : Génération de questions, Compréhension de documents, Question/Réponse, Humanités numériques.

KEYWORDS: Question Generation, Machine reading Question Answering, Digital Humanities.

/!\ Cet article est un résumé/condensé de deux publications faites à ce sujets : /!

- Question Generation and Answering for exploring Digital Humanities collections (LREC 2022) (Bechet *et al.*, 2022)
- Génération de questions à partir d'analyse sémantique pour l'adaptation non supervisée de modèles de compréhension de documents (TALN 2022) (Antoine *et al.*, 2022)

Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain *(LIFT)* et Traitement Automatique des Langues *(TAL)*,
Marseille, France, 14 au 15 novembre 2022 , pages 119–122,

1 Introduction

Machine Reading Comprehension and Question Generation are two *mirror* tasks of Natural Language Processing (NLP). Traditionally handled by complex *pipelines* based on very different models, information retrieval for question answering models and parsing for question generation, they have recently been unified since the advent of *end-to-end* methods based on pre-trained language models.

Thus, as presented in (Du *et al.*, 2017), question generation can be modeled as a text generation task where a sequence-to-sequence model is trained to *translate* a sequence of words representing a sentence or a passage into another sequence of words representing a question, without going through an explicit linguistic analysis as it used to be the case. On the other hand, automatic document understanding can be seen as a labeling task consisting in learning, from a pair *(text,question)*, which words should be labeled with the labels *start of answer* and *end of answer* in the text.

The development of pre-trained language models used in conjunction with large corpora of triplets *question/answer/context* such as *SQuAD* (Rajpurkar *et al.*, 2016) can be used to directly train both translation models *answer-to-question* and labeling models *question-to-answer*. While the performance of these models is impressive on these reference corpora, generally containing text from Wikipedia and simple questions obtained by crowdsourcing, the generalization of these models to corpora containing more complex texts and less literal questions remains a challenge. It is in this context that we propose in this study a method for unsupervised adaptation of text comprehension models based on automatic question generation.

The contributions of this study are at two levels : on the one hand, the comparison of different methods of encoding semantic information for the generation of questions evaluated with respect to their capacity to train a question/answer model on a new corpus of texts ; on the other hand, the study of the generalization capacity of comprehension and question generation models on a new corpus containing texts and questions more complex than those that can be found in reference corpora such as *SQuAD* (Rajpurkar *et al.*, 2016) for English or *FQuAD* (d'Hoffschmidt *et al.*, 2020) for French.

In order to study how the current boost in performance in NLU models on benchmark data translates to real-life settings, the applicative framework considered here is the exploration of digitized collections by professional users that are used to analyze archives in order to perform Social Science research. We chose to focus on the question/answering paradigm, as asking questions and looking for answers is at the same time a natural way for researchers to explore archives and also the task that received the most attention in recent language understanding studies, especially since the release of large training data such as SQuAD (Rajpurkar *et al.*, 2016) [1].

In this paper we will present first the *self-management* corpus, a collection of a French journal ranging over 20 years from 1966 to 1986, which has been chosen as our archival material in the project, then we will highlight the differences between benchmark corpora usually based on *Wikipedia* and digitized archive collections. We will then present the question/answering paradigm, the annotation scheme developed in Archival and point out the differences between the kinds of questions that can be made by professional users and those used in Machine Reading datasets such as *SQuAD*. We will describe the question generation and question answering models that have been developed to adapt a Machine Reading model trained on Wikipedia to the *self-management* corpus of the project without any supervision. Finally we will present the first results obtained on the *self-management* dataset with this adapted Machine Reading model.

1. ttps://rajpurkar.github.io/SQuAD-explorer/

2 The self-management corpus

2.1 Origin of the collection

The "self-management" notion falls within the large spectrum of social sciences. It concerns daily social environment, economic life, as well as political life, education, ecology, culture, architecture, ... It addresses populations structure, the relationship of populations with resources, the political, legal and administrative framework of society and the authority relations between individuals and groups.

Since the 1960's, the FMSH [2] foundation's library has gathered a pluridisciplinary multilingual mixed collection (archives and documents) about self-management (*autogestion* in French). It gathers around 25000 pieces : books, journals, reports, leaflets, correspondences.

2.2 Corpus description

For this study, we are particularly interested in the *Autogestion* journal [3] which is distributed in its digitized form by the French Persée organization. We are using a version of the corpus that has been OCRized with Tesseract without manual corrections. Hence data are not free of OCR errors but the structure of the journal (mono-column, few figures) implies that the OCR quality is good (further studies could imply precise evaluation of OCR quality and impact of OCR errors on downstream NLP tasks but for this study, OCR output are taken *as is*).

The resulting corpus is composed of 46 issues ranging over 20 years, for an overall amount of 6298 pages and 1.98M tokens.

2.3 Specificities of texts from an NLP point of view

Most studies in Information Extraction or Question Answering are carried out on Wikipedia pages. Wikipedia documents are particularly well suited for theses tasks as they intrinsically dedicated to convey factual information. Another characteristic of Wikipedia is that articles are supposed to follow a Neutral Point of View policy [4]. Recent work (Bertsch & Bethard, 2021) aims at detecting so-called puffery (*i.e* sentences that do not respect that policy, which are tagged by editors as "peacock phrases") but this phenomenon remains very rare. On the contrary, texts that are relevant for Digital Humanities and studies related to Social Science are not only factual and neutral documents but also essays or articles that reflect the writer's point of view. Description of events are not only depicted by facts but with deeper analysis of the previous notions or influences that yielded this event as well as their consequences and how they influenced the thinking of other actors.

Références

ANTOINE E., AUGUSTE J., BECHET F. & DAMNATI G. (2022). Génération de question à partir

2. Fondation Maison des Sciences de l'Homme, https://www.fmsh.fr/
3. https://www.persee.fr/collection/autog
4. https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view

d'analyse sémantique pour l'adaptation non supervisée de modèles de compréhension de documents. In Y. ESTÈVE, T. JIMÉNEZ, T. PARCOLLET & M. ZANON BOITO, Éds., *Traitement Automatique des Langues Naturelles*, p. 104–115, Avignon, France : ATALA. HAL : hal-03701494.

BECHET F., ANTOINE E., AUGUSTE J. & DAMNATI G. (2022). Question Generation and Answering for exploring Digital Humanities collections. In *13th Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France. HAL : hal-03719368.

BERTSCH A. & BETHARD S. (2021). Detection of puffery on the english wikipedia. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, p. 329–333.

D'HOFFSCHMIDT M., BELBLIDIA W., HEINRICH Q., BRENDLÉ T. & VIDAL M. (2020). FQuAD : French question answering dataset. In *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 1193–1208, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.107.

DU X., SHAO J. & CARDIE C. (2017). Learning to ask : Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1342–1352.

RAJPURKAR P., ZHANG J., LOPYREV K. & LIANG P. (2016). Squad : 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 2383–2392 : Association for Computational Linguistics. DOI : 10.18653/v1/D16-1264.

HOLINET: Holistic Knowledge Graph for French
Graphe de connaissances holistique pour le français

Jean-Philippe Prost[1]
(1) Laboratoire Parole et Language, CNRS – Aix-Marseille Université, France
Jean-Philippe.Prost@univ-amu.fr

RÉSUMÉ

HOLINET est un graphe de connaissances pour le français, qui vise à fournir une perspective holistique de la représentation des connaissances linguistiques. Ainsi, il approche la langue à la fois comme tout, et comme la somme de ses parties sur les différentes dimensions linguistiques. Nous formulons l'hypothèse qu'une telle modélisation holistique des connaissances linguistiques facilitera le traitement automatique du language et améliorera les performances des applications en aval. HOLINET ouvre de nouvelles pistes de recherche en tant que graphe de connaissances qui intègre des connaissances syntaxiques de référence et des connaissances lexico-sémantiques, et qui combine constituance et dépendance. L'encodage de HOLINET en un modèle de plongement de graphe de connaissances reste une perspective saillante à explorer.

ABSTRACT

HOLINET : Holistic Knowledge Graph for French

HOLINET is a knowledge graph (KG) for French, which aims to provide a holistic perspective on language knowledge representation. As such, it approaches language as a whole, as well as a sum of its parts on various linguistic dimensions. We hypothesize that a holistic modelling of language knowledge will ease its processing and improve the performance of downstream applications. HOLINET opens new avenues of research as a KG which integrates gold-standard syntactic knowledge along with lexical semantic one, and which is open to combining constituency and dependency information. The computation of a KG embedding model, for instance, is a salient option to investigate.

MOTS-CLÉS : Graphe de connaissances, réseau lexico-sémantique, grammaire syntagmatique.

KEYWORDS: Knowledge Graph, Lexical-Semantic Network, Phrase Structure Grammar.

1 Introduction

Knowledge Graphs (KGs) have become a corner stone of modern Artificial Intelligence (AI), for the central role they play in a variety of downstream applications, such as QA, recommender systems, semantic parsing, etc. They suit both symbolic and sub-symbolic types of processing, since they may be involved in these applications in plain form, as part of graph-theoretical processes (e.g. path-based reasoning), or in sub-symbolic form, where the graph is converted to a numerical representation, such as knowledge graph embeddings.

As far as NLP applications are concerned, KGs are often relied on for lexical and semantic knowledge, usually through KG emmbeddings, while syntax is usually gathered from other sources, typically

Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain *(LIFT) et* Traitement Automatique des Langues *(TAL),*
Marseille, France, 14 au 15 novembre 2022 , pages 123–130,

annotated corpora. In this case, the integration between syntax and lexical semantics is done by the application. But what if the integration was done earlier, in the KG ? Would embeddings encoded on such a KG perform better than the existing models for the relevant NLP tasks (e.g. semantic parsing) ?

More generally, while the pipeline software architecture, which steps from one linguistic dimension to the next, has been typical for decades for most NLP applications, it often prevents many potential interactions across dimensions from actually occurring. A variety of sentence-level ambiguities, for instance, require the full sentence to be parsed morphologically, then syntactically, then semantically, prior to being disambiguated through a pipeline. Knowledge graphs provide a convenient means for heterogeneous knowledge to interact rather seamlessly.

Prior to addressing questions such as the performance of syntax-semantic Knowledge Graph embeddings in NLP tasks, this work focuses on the construction of such an integrated KG, and the problems that come along with it. Section 2 introduces some background knowledge and review the literature. Section 3 presents the graph model underpinning the HOLINET knowledge graph and its implementation, along with evaluation figures. Section 4 discusses further works, and section 5 concludes.

2 Background and literature review

What is a Knowledge Graph ? As Hogan *et al.* (2021, p. 2) put it, "[t]he definition of a "*knowledge graph*" remains contentious". Their own definition is an inclusive one,

> "(. . .) where we view a knowledge graph as a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between these entities. The graph of data (aka data graph) conforms to a graph-based data model, which may be a directed edge-labelled graph, a property graph, etc. (. . .). By knowledge, we refer to something that is known".

Literature review Knowledge Bases for natural language, whether structured as graphs or not, are rarely holistic, in the sense that they would merge, and present, all linguistic dimensions as a whole, within a single and homogeneous structure. The Linguistic Linked Open Data (LLOD) initiative links together different resources. LLOD's interest goes towards representation format problems, or federation of multiple data sources, or interoperability. But the seamless integration of heterogeneous data is not its prime concern – especially not so at the interface between syntax and lexical semantics. In fact, as far as we know, LLOD only links syntactic resources through annotated corpora. We do not know of any grammatical knowledge base, regardless of the language. Faralli *et al.* (2020) is concerned with integrating 5 resources already linked through LLOD : ConceptNet (Speer *et al.*, 2017), DBpedia (Lehmann *et al.*, 2015), WebIsAGraph (Faralli *et al.*, 2019), WordNet (Miller, 1995) and the Wikipedia category hierarchy. No syntactic resource involved.

The lexico-semantic graphs relie on various structures. WordNet is a network of 150K+ words, organised in 170 000 synsets, which can be seen as concepts. WOLF (Sagot & Fier, 2008) is a French version of it. The French Lexical Network (*Réseau Lexical du Français*) (RLF) from Polguère (2014) relies on the notion of *lexical function* as defined by Igor Mel'čuk. JeuxDeMots (JDM) is another lexico-semantic network for French (Lafourcade, 2007), which implements the same notion and generalises it beyond lexical knowledge. JDM is made up of 16,5+ millions nodes, including

5,2+ millions terms, 400+ millions relationships and 150+ relationship types. The words are terms, concepts and symbolic information. The relationship types are lexical, morphological, pragmatical, logical, ontological, etc.

None of these resources include grammatical knowledge. FrameNet (Baker *et al.*, 1998), which is dedicated to *frame semantics* (Fillmore, 2008), relates the semantic frames with each others through semantic relationships to constitute a network. Each frame is illustrated with prototypical utterances, annotated with syntax. However, the syntactic analysis is expressed through text annotations, and is not, strictly speaking, integrated in the network.

The literature shows that syntactic knowledge is mainly represented through annotated resources, and to some extent through symbolic grammars, such as the HPSH grammars (Pollard & Sag, 1994) from the DELPH-IN consortium (Copestake & Flickinger, 2000), or meta-grammars such as FRMG (de La Clergerie, 2005).

3 What is HOLINET about ?

HOLINET aims to provide a holistic perspective on language knowledge representation. Holistic in that all linguistic dimensions, although heterogeneous by nature, are integrated within the same data structure. As such, it approaches language as a whole, as well as a sum of its parts on various dimensions. One of the main motivation for such an approach is to overcome some of the issues raised by the traditional pipeline architecture for NLP applications.

The HOLINET graph model has already been thoroughly detailed in (Prost, 2022), along with its automated construction process [1] [2]. The resource is original in that it integrates lexical semantic knowledge and grammar knowledge within a single graph. The lexical semantic layer is the lexico-semantic network JeuxDeMots (JDM) (Lafourcade, 2007). It conveniently represents the POS categories of all the terms in the network. conveniently, because the POS categories will serve as the interface between JDM and the grammar layer to come. The grammar layer is made up of a phrase structure grammar, which may be seen as a set of context-free rewrite rules, such as Rule (1) illustrated in Figure 1. The POS in a rule which pre-exist in JDM are as many anchor nodes. The phrasal categories, e.g. Noun Phrase (NP) or Adjective Phrase (AP), do not pre-exist in JDM, hence are created for the grammar layer in HOLINET. Note that the type `n_pos` in HOLINET generalises both the actual POS categories and the phrasal categories.

Figure 1 illustrates the graph model of the grammar layer for the example rule (1). The POS categories are pre-existing nodes of the JDM type `n_pos`. The terms are related to their respective POS categories with pre-existing relationships of the JDM type `r_pos`. Every rewrite rule is reified as a node, typed `n_g_cfRule`. The left-hand side of each rule is itself reified as a POS node (of type `n_pos`), and every rule is connected to its left-hand side POS node with the `r_g_rewrites` relation. Meanwhile, on the right-hand side (RHS) of each rule, every constituent POS is connected to its rule with the `r_g_constitutes` relation. In order to allow redundancy of POS, like here the AP, every constituent on the RHS is related to its POS node with an `r_g_instantiates` relation. The feature structure next to Rule (1) details the properties associated with the `n_g_cfRule` node

1. HOLINET v1.0 is distributed by ORTOLANG (`https://hdl.handle.net/11403/holinet-1-0/v1`) under a Creative Commons Attribution 4.0 International licence (CC-BY 4.0).
2. All the software involved in the creation process is publicly available as git repositories on sourceforge.net. See (Prost, 2022) for more details.

(1) NP →DET#fsi NC#fsc AP AP

$$\text{RULE (1)}\left[\begin{array}{ll}\text{TYPE} & \texttt{n\_g\_cfRule}\\ \text{RULE} & \left[\begin{array}{ll}\text{LHS} & [\text{CAT}\ NP]\\ \text{RHS} & \left\langle \underset{\text{DET\#FSI}}{\left[\begin{array}{ll}\text{CAT} & DET\\ \text{CPOS} & D\\ \text{GEND} & f\\ \text{NUM} & s\\ \text{SUBCAT} & ind\end{array}\right]}, \underset{\text{N\#FSC}}{\left[\begin{array}{ll}\text{CAT} & NC\\ \text{CPOS} & N\\ \text{GEND} & f\\ \text{NUM} & s\\ \text{SUBCAT} & c\end{array}\right]}, [\text{CAT}\ AP], [\text{CAT}\ AP]\right\rangle\end{array}\right]\end{array}\right]$$

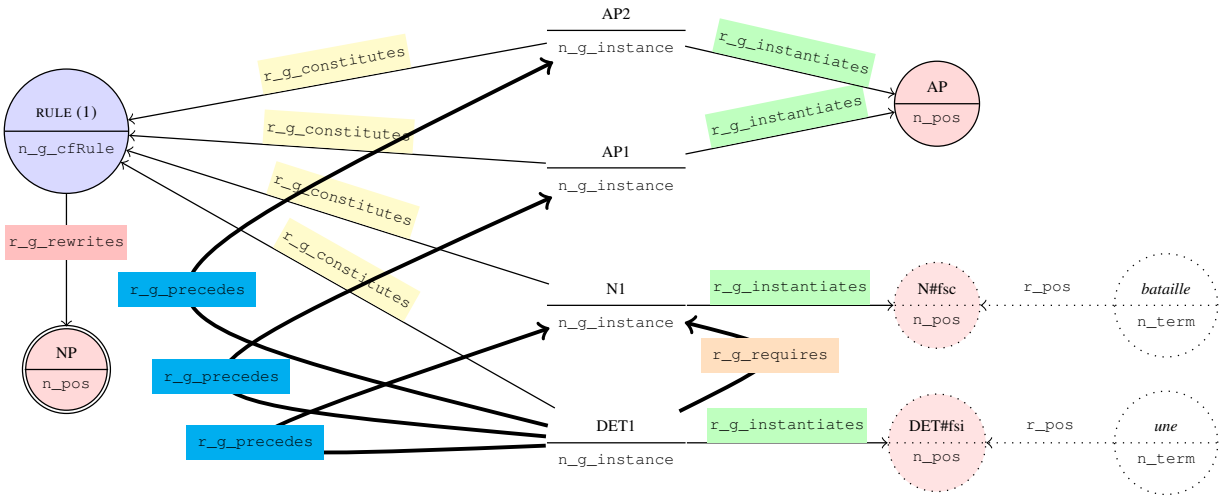


FIGURE 1 – HOLINET Graph model for the example context-free grammar rewrite rule (1).
The dotted nodes and edges originate in JDM. They serve as anchorage for connecting the grammar layer to JDM.

labelled RULE (1).

Next to the Immediate Dominance relationship, represented by the `r_g_constitutes` relations, the model also represents other relationships (in bold), such as Linear Precedence (`r_g_precedes`), and Requirement (`r_g_requires`). Their semantics is borrowed to the corresponding well-defined *properties* in Property Grammar (PG) (Blache, 2001), a formal framework for specifying constraint-based grammars. Further works will investigate the integration of even more relationships, such as *obligation* (to model a phrase's head), *agreement* and *dependency* (for dependency grammars).

The creation process in short For the sake of generalisation, we assume (a) a constituency treebank, and (b) a lexical network. The lexical entries in the network are expected to be related to their POS categories, where the relationship is of type `r_pos`, and the POS are as many nodes. If necessary, we assume a mapping between the two POS tagsets, i.e. the network's and the treebank's. The creation process is, then, made up of the following steps :

1. *read/extract* the implicit CFG from the constituency trees in the treebank
2. *derive* the additional relationships from the CFG (linear precedence, requirement, etc.)
3. *assess* the truth values to be assigned to the relationships (default is all true).
   Given a corpus CFG, the process of derivation (step 2), then assessment (step 3) of the relationships we are interested in, is equivalent to the so-called "compilation process" of a Property Grammar, as described by Prost *et al.* (2016).
4. *map* the treebank tagset to the network's (or the other way around)

5. *create* the required grammar triples (i.e., nodes and relations), according to the augmented CFG (i.e., with the additional relationships from step 2)

6. *merge* the two layers and check their consistency (detailing this step further goes beyond the scope of this paper).

Experiments and first evaluation The creation process has been implemented with the version 1.0 (2016) of the FTB annotated in the Penn Treebank format, and the version of JDM extracted from the dump dated 01/11/2021. But we believe that the process is trivial enough to be easily adapted to other resources.

Since we are primarily concerned with integrating the grammar layer with the lexical-semantic layer, and since the anchorage of the grammar layer to JDM is achieved through the POS nodes shared by the two layers, we want to measure the proportion of the POS categories required by the grammar layer (i.e., found in the FTB) that are actually found in JDM. We are only interested in the actual POS, not the phrase categories, in spite of the fact that both are modelled in HOLINET as nodes of type `n_pos`. Our evaluation is reported in Table 1. It shows that only 30.1% of the POS categories

	Connected	Disconnected	Null	**Total**
Num. nodes	22,742	43361	9,408	**75511**
%	30.1%	57.4%	12.5%	**100**

TABLE 1 – proportion of the POS categories found in FTB grammar rules, that are actually found in JDM (connected), or not (disconnected). The phrase categories, although typed `n_pos`, have been taken out of the picture. The 'Null' value stands for mapping issues.

involved in the grammar are actually found *as such* in JDM.

Most of the discrepancies between the two layers come from the choice JDM makes to split among several nodes the different attributes associated with a category (e.g. gender, number, etc.), while on the grammar layer a single node is required. For instance, the FTB tag `P+PRO##cpos=P+PRO|g=f|n=p|p=3|s=rel##` is, theoretically, mapped to the JDM label `Pre+Pro:Fem+PL+Rel`. The expected corresponding labelled node is actually absent from JDM, but the following nodes typed `n_pos` are present : `Pre:`, `Pro:Fem+PL`, `Pro:Rel`, and `Pre:`. That is, all the required information is present, but split across distinct nodes.

Note that coming up with an exact algorithm, which would create the required merged nodes from the distinct ones is not trivial. Take, for instance, the term 'les'. It is connected in JDM to the following distinct `n_pos` nodes : `Det:`, `Pro:`, `Pro:Pers:COD`, `Pro:Pers`, `Pro:PL+P3`, `Det:Fem+PL`, `Det:Mas+PL`, `Det:InvGen+PL`, `Gender:Mas`, `Number:Plur`, `Gender:Fem`, `Defini:`. Quite obviously computing all the combinations is an option that would not make much sense. Improving the mapping between JDM and the HOLINET grammar layer is, therefore, ground for further investigation.

4 Applications and further works

Would a KG embedding model computed from HOLINET with both semantic and syntactic relationships improve downstream applications, such as semantic parsing ? The injection of syntactic knowledge in neural models for semantic parsing, whether deep or shallow, has been consistenly shown to improve performance. Roth & Lapata (2016) use a dependency path embedding model to improve a Recurrent Neural Network model for Semantic Role Labelling (SRL). Wang *et al.* (2019) show that the injection of syntax as input features into three different neural SRL encoders significantly improves performance. Their works also show that constituency features perform best, ahead of dependency and categorical constituency spans. Xu *et al.* (2018) combine word order, dependency and constituency features within graph embeddings. More recent works by Fei *et al.* (2021) succesfully investigate the combination of constituency and dependency through TreeLSTM and Graph Convolutional Network. Kurtz *et al.* (2019) suggest that only gold-standard syntactic information, as opposed to automatically predicted one, improves the performance of a deep neural architecture for semantic parsing. Moreover, the integration of syntactic knowledge along with lexical and semantic knowledge within the same embeddings is modern ground for investigation (Limisiewicz & Mareček, 2020; Al-Ghezi & Kurimo, 2020). In line with this body of work, HOLINET opens new avenues of research as a KG which integrates gold-standard syntactic knowledge along with lexical semantic one, and which is open to combining constituency and dependency information. The computation of a KG embedding model is, then, a salient option to investigate.

More avenues of research

— Could HOLINET integrate other types of grammar knowledge, such as dependency grammar, or Construction Grammar, and how ?

— Could the interaction between syntactic and semantic knowledge be captured in deductive and/or inductive reasoning processes for link prediction ? For desambiguation ?

5 Conclusion

In this paper we investigated the question of the integration of grammar knowledge and lexical semantic knowledge within a homogeneous graph structure, in order to construct a holistic knowledge graph for French. Our motivation is to implement an environment that enables the investigation of integrated syntax-semantic knowledge graph embeddings and their performance in downstream applications, or graph-theoretical algorithms for automated reasoning.

We presented a graph model for a phrase structure grammar, and we showed how to merge it with a lexical semantic network through a shared tagset for POS categories. We experimented the creation procedure with the French treebank (FTB) annotated for constituency, and the lexico-semantic network JeuxDeMots (JDM). Our evaluation shows that 30.1% of the POS required by the FTB can actually be found in JDM as a single node. This figure does not jeopardize the graph model as such, but rather shows that, although all the required information can be found in JDM, further work is still necessary in order to better map the annotation schemes.

Références

AL-GHEZI R. & KURIMO M. (2020). Graph-based syntactic word embeddings. In *Proceedings of the Graph-based Methods for Natural Language Processing (TextGraphs)*, p. 72–78.

BAKER C. F., FILLMORE C. J. & LOWE J. B. (1998). The Berkeley FrameNet Project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, volume 1, p. 86–90.

BLACHE P. (2001). *Les Grammaires de Propriétés : des contraintes pour le traitement automatique des langues naturelles*. Hermès Sciences.

COPESTAKE A. & FLICKINGER D. (2000). An open source grammar development environment and broad-coverage English grammar using HPSG.

DE LA CLERGERIE É. (2005). From Metagrammars to Factorized TAG/TIG Parsers. In *Proceedings of IWPT'05 (poster)*, Vancouver, Canada.

FARALLI S., FINOCCHI I., PONZETTO S. P. & VELARDI P. (2019). Webisagraph : A very large hypernymy graph from a web corpus. In *Proceedings of the Sixth Italian Conference on Computational Linguistics*, p. 13–15, Bari, Italy.

FARALLI S., VELARDI P. & YUSIFLI F. (2020). Multiple knowledge graphdb (mkgdb). In *Proceedings of The 12th Language Resources and Evaluation Conference*, p. 2325–2331.

FEI H., WU S., REN Y., LI F. & JI D. (2021). Better combine them together ! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, p. 549–559 : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-acl.49.

FILLMORE C. J. (2008). *Cognitive Linguistics : Basic Readings*, chapitre Frame semantics, p. 373–400. De Gruyter Mouton. DOI : doi :10.1515/9783110199901.373.

HOGAN A., BLOMQVIST E., COCHEZ M., D'AMATO C., MELO G. D., GUTIERREZ C., KIRRANE S., GAYO J. E. L., NAVIGLI R., NEUMAIER S. *et al.* (2021). Knowledge graphs. *ACM Computing Surveys (CSUR)*, **54**(4), 1–37.

KURTZ R., ROXBO D. & KUHLMANN M. (2019). Improving semantic dependency parsing with syntactic features. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, p. 12–21, Turku, Finland : Linköping University Electronic Press.

LAFOURCADE M. (2007). Making people play for Lexical Acquisition. In *Proc. SNLP 2007*, p. 13–15, Pattaya Thailande, December. 8 p : 7th Symposium on Natural Language Processing.

LEHMANN J., ISELE R., JAKOB M., JENTZSCH A., KONTOKOSTAS D., MENDES P. N., HELLMANN S., MORSEY M., VAN KLEEF P., AUER S. *et al.* (2015). Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, **6**(2), 167–195. DOI : 10.3233/SW-140134.

LIMISIEWICZ T. & MAREČEK D. (2020). Syntax Representation in Word Embeddings and Neural Networks–A Survey. arXiv preprint arXiv :2010.01063.

MILLER G. A. (1995). WordNet : a lexical database for English. *Communications of the ACM*, **38**(11), 39–41.

POLGUÈRE A. (2014). From Writing Dictionaries to Weaving Lexical Networks. *International Journal of Lexicography*, **27**(4), 396–418.

POLLARD C. & SAG I. (1994). *Head-driven Phrase Structure Grammar*. University of Chicago Press.

PROST J.-P. (2022). Integrating a phrase structure corpus grammar and a lexical-semantic network : the holinet knowledge graph. In *Proceedings of LREC 2022 the 13th Language Resources and Evaluation Conference*, p. 613–622, Marseille, France : European Language Resources Association European Language Resources Association. HAL : hal-03655636.

PROST J.-P., COLETTA R. & LECOUTRE C. (2016). Compilation de grammaire de propriétés pour l'analyse syntaxique par optimisation de contraintes. In *Actes de TALN 2016, 23ème conférence sur le Traitement Automatique des Langues Naturelles*, p. 396–402, Paris, France : Association pour le Traitement Automatique des Langues.

ROTH M. & LAPATA M. (2016). Neural semantic role labeling with dependency path embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1192–1202 : Association for Computational Linguistics. DOI : 10.18653/v1/P16-1113.

SAGOT B. & FIER D. (2008). Construction d'un WordNet libre du français à partir de ressources multilingues. In *Proceedings of TALN 2008*, Avignon, France.

SPEER R., CHIN J. & HAVASI C. (2017). Conceptnet 5.5 : An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

WANG Y., JOHNSON M., WAN S., SUN Y. & WANG W. (2019). How to best use syntax in semantic role labelling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 5338–5343 : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1529.

XU K., WU L., WANG Z., YU M., CHEN L. & SHEININ V. (2018). Exploiting rich syntactic information for semantic parsing with graph-to-sequence model. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 918–924 : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1110.

Langue des Signes Française : État des lieux des ressources linguistiques et des traitements automatiques

Annelies Braffort
Université Paris-Saclay, CNRS, LISN, Campus bât 507 - Rue du Belvédère, 91405 Orsay, France
annelies.braffort@lisn.upsaclay.fr

RÉSUMÉ

Cet article présente un état des lieux sur les ressources disponibles pour la recherche sur la Langue des Signes Française (LSF), ainsi que sur son traitement automatique. Après une mise en contexte sur les langues des signes et plus particulièrement la LSF, l'article recense les ressources disponibles pour la recherche sur la LSF en général, puis plus précisément sur leur utilisation pour les recherches en traitement automatique des langues des signes, en s'appuyant sur des exemples de projets représentatifs récents ou en cours.

ABSTRACT

French Sign Language : Overview of language resources and automatic processing

This article presents an overview of the resources available for research on French Sign Language (LSF), as well as on its automatic processing. After a background on sign languages and more specifically on LSF, The article lists the resources available for research on LSF in general, and then more specifically on their use for research in automatic sign language processing, based on examples of recent or ongoing representative projects.

MOTS-CLÉS : Langue des Signes Française, LSF, Corpus, Traitement Automatique.

KEYWORDS: French Sign Language, LSF, Corpus, Automatic Processing.

1 Les langues des signes : langues naturelles visuo-gestuelles

Les langues des signes (LS) sont des langues naturelles pratiquées au sein des communautés de Sourds [1]. Les LS sont des langues sans écriture, relevant, à ce titre, de la modalité orale, du face-à-face, même transmis en différé. Elles sont visuo-gestuelles, produites par de nombreux articulateurs corporels, les mains et les bras bien sûr, mais aussi le visage, les épaules, la tête, le regard et plusieurs composantes faciales, qui peuvent être activés plus ou moins *simultanément* et perçues par les yeux. Cette modalité de perception est bien adaptée à l'interprétation de mouvements organisés dans l'espace. De fait, cet espace joue un rôle fondamental dans la structuration du discours et est nommé *espace de signation*. Les personnes nées sourdes, qui ont une expérience du monde basée sur la perception visuelle et non sonore, expriment avoir un mode de pensée visuel qui leur fait privilégier des constructions linguistiques qui permettent de dire mais aussi de *montrer*.

Ainsi, les LS comportent des constructions linguistiques de nature diverse. Certaines d'entre elles sont considérées comme des unités spéciales appelées *signes lexicaux* ou simplement *signes*, qu'on peut

1. "Sourd" avec un "S" majuscule désigne une identité culturelle, historique et linguistique.

Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain *(LIFT) et* Traitement Automatique des Langues *(TAL)*,
Marseille, France, 14 au 15 novembre 2022 , pages 131–138, ©2022 CNRS

lister dans un dictionnaire. D'autres types de constructions, très illustratives, exploitent les spécificités mentionnées ci-dessus et peuvent être présentes à hauteur de 20 à 80% selon le type de discours (Sallandre *et al.*, 2019).

Les LS se sont développées au fil du temps et sont très liées à la culture. Tout comme les langues parlées, elles possèdent de multiples formes, des dialectes et des variations locales. L'édition courante de l'Ethnologue [2] répertorie 157 LS, mais il en existe cerftainement d'autres qui n'ont pas encore été documentées ni identifiées. Les différences se situent surtout au niveau du lexique, mais il existe des similitudes sur le plan grammatical car toutes les LS exploitent les capacités de multilinéarité, d'utilisation de l'espace et d'iconicité.

La langue des signes française (LSF) est pratiquée en France et dans la partie francophone de la Suisse. Le nombre de locuteurs de LSF, qui peuvent être des personnes sourdes ou entendantes (les membres de la famille et les proches de personnes Sourdes) n'est pas connu avec précision. On cite souvent un nombre variant de 100 000 à 300 000 locuteurs. Il existe également des langues des signes tactiles utilisées par les personnes avec surdicécité. Elles diffèrent significativement des LS visuelles en ce que des éléments comme l'expression faciale sont remplacés par des informations tactiles.

La présence de la LSF dans les médias a augmenté ces dernières années, notamment depuis l'adoption de la loi de 2005 qui la reconnaît comme langue à part entière [3]. Malgré cette loi qui impose aux établissements qui reçoivent du public de rendre accessible les informations quel que soit le type de handicap, encore très peu d'informations sont disponibles en LSF. Cela pose des problèmes d'accessibilité aux informations pour les personnes Sourdes qui peuvent avoir une maîtrise limitée du français, même écrit, puisque c'est souvent pour eux une langue seconde. Des outils tels que la traduction automatique ou du moins l'aide à la traduction, du texte vers la LSF pour améliorer l'accessibilité, mais aussi dans l'autre sens, pour le sous-titrage de vidéos de LSF, ainsi que l'outillage des vidéos de LS en général, seraient d'une grande utilité pour ces langues et la communauté concernée.

Cet article présente un état des lieux sur les ressources disponibles pour la recherche sur la LSF en général, puis plus précisément sur leur utilisation pour les recherches en traitement automatique des langues des signes.

2 Les corpus de LSF : peu nombreux et de petite taille

Les corpus de LSF répertoriés sur les plateformes d'archivage et de diffusion de corpus tels que Cocoon [4] ou Ortolang [5] sont peu nombreux. On dénombre à l'heure actuelle 22 entrées pour la LSF (il existe aussi des dépôts relatifs à d'autres LS). La plupart des dépôts sur Cocoon sont des numérisations de conférences ou séminaires datant des années 1990 et donc à visée patrimoniale. On y trouve aussi LS-COLIN, le premier corpus de LSF enregistré en studio spécifiquement pour la recherche en 2002. Les dépôts sur Ortolang ont tous été créés dans le cadre de projets de recherche. Presque tous sont des corpus de laboratoire, c'est-à-dire enregistrés en studio avec un objectif de recherche.

Le tableau 1 recense uniquement les corpus de LSF conçus pour la recherche et disponibles actuelle-

2. https://www.ethnologue.com/subgroups/sign-language
3. https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000809647/
4. https://cocoon.huma-num.fr/
5. https://www.ortolang.fr/

ment sur les plateformes Ortolang et Cocoon. Ils sont de nature très variée selon le type de contenu (production très contrôlée, monologue, dialogue, ou encore repas en famille), les lieux d'enregistrement (en studio, chez des particuliers, dans la rue) ou les conditions techniques (de 1 à 5 caméras synchronisées, système de capture de mouvement). Certains corpus ne comportent que des données primaires et d'autres sont traduits ou sous-titrés en français et annotés. Là encore la nature et le format des annotations est très variée, depuis de simples gloses pour identifier les signes lexicaux jusqu'à des descriptions fines des constructions linguistiques ou du mouvement des articulateurs. La plupart des corpus sont de très petite taille (moins de 10h), à l'exception de CREAGEST et MEDIAPI-SKEL.

CREAGEST [6] est un corpus de laboratoire dont la partie déposée sur Ortolang comporte 156h de LSF produites par plus de 82 signeurs. Il a été créé avec pour objectif la constitution et la documentation de vidéos de productions gestuelles incluant des productions en LSF d'enfants et d'adultes sourds et des productions de gestualité naturelle pour des études en linguistique et en acquisition (Garcia *et al.*, 2013). Il est a l'heure actuelle le plus gros corpus de LSF, mais à ce jour seule une partie des données primaires a été déposée sur Ortolang.

MEDIAPI-SKEL [7] a été élaboré à partir de contenus fournis par Média'Pi !, un média en ligne bilingue en LSF et en français. De nombreux thèmes y sont abordés, sous forme d'actualités présentées par des journalistes ou présentateurs sourds, d'interviews ou reportages avec plusieurs intervenants, ou d'autres formes plus atypiques (photos-reportages, BDs). La langue première de ces contenus est la LSF. Dans un second temps, un sous-titrage en français est produit. La version actuelle déposée sur Ortolang contient 27 heures de données préparées pour des études en TALS (Bull *et al.*, 2020) : des sous-titres alignés avec des vidéos comportant une représentation simplifiée des signeurs sous forme de squelettes 2D avec des points clés sur le visage, les mains et le corps (figure 1). Les vidéos originales sont accessibles par le biais d'un abonnement auprès de Média'Pi.

FIGURE 1 – Extrait du corpus MEDIAPI-SKEL

Ces deux exemples illustrent bien la grande diversité de ces corpus, élaborés dans le cadre de projets avec un objectif précis. Si certains pourraient être utilisés dans d'autres contextes que ceux prévus initialement, dans la pratique cela semble peu courant, sauf de rares cas où le corpus a été prévu dès le départ pour permettre des études pluridisciplinaires, tel que le corpus MOCAP1 qui a été utilisé pour des études en sciences du mouvement, en linguistique et en informatique (Benchiheub *et al.*, 2016; Collomb *et al.*, 2018; Bigand, 2021).

6. https://www.ortolang.fr/market/corpora/ortolang-000926/
7. https://www.ortolang.fr/market/corpora/mediapi-skel/

Ainsi à l'heure actuelle, pour la LSF, les corpus disponibles pour la recherche sont encore peu nombreux et de taille très limitée. Au niveau international, certaines LS ont pu bénéficier de financements conséquents qui ont permis la création de corpus de grande taille (pour les LS) et bien documentés, comme par exemple le corpus de LS allemande *DGS Corpus* [8] de 560h, qui a été élaboré dans le cadre d'un projet financé sur 15 ans. Deux parties de plus de 50h sont accessibles en ligne : *My DGS*, accessible à tout public et fourni avec des sous-titrages et *My DGS annotated*, à destination des chercheurs. Un tel corpus nécessite le développement d'outils permettant son exploitation et donc des études en traitement automatique des LS (TALS).

La section suivante dresse un état des lieux des études actuelles en TALS sur la LSF, en lien avec les corpus existants.

3 Corpus pour le traitement automatique

La recherche en traitement automatique des LS est beaucoup plus récente que celle dédiée aux langues parlées ou écrites. Elle est particulièrement active dans le domaine de la reconnaissance automatique, mais il existe aussi des projets en génération et depuis peu en traduction automatiques.

Dans le cadre du projet Européen en cours EASIER [9] sur la traduction automatique entre certaines langues écrites et langues des signes d'Europe, un livrable [10] recense les ressources linguistiques qui peuvent être utilisées pour le TALS. Il répertorie en particulier les corpus de LS européennes de grande taille (pour les LS) qui peuvent être utilisés comme données d'entraînement de haute qualité pour la traduction automatique.

On peut distinguer deux types de ressources : les corpus de recherche et les données télédiffusées. Les corpus de recherche, et plus particulièrement ceux créés en vue d'études en linguistique, offrent des données de qualité élevée accompagnés d'une transcription et d'une annotation linguistique, mais ils n'en existent pas de grande taille pour toutes les LS et ils sont malgré tout de taille réduite par rapport aux besoins pour les approches à base d'apprentissage. Les données télédiffusées sont souvent disponibles en quantité relativement importante et comporte généralement des sous-titres synchronisés avec la parole. Cependant il s'agit la plupart du temps de LS produite en direct par un interprète, et soumise aux contraintes temporelles du direct et de la structure du discours oral qui est interprété. De plus, la qualité des alignements entre la LSF et les sous-titres peut être assez mauvaise car les sous-titres ne sont pas alignés avec la LS et il peut y avoir plusieurs secondes de décalage.

Un nouveau type de ressource consiste en des données télédiffusées de LS non interprétées produites par des locuteurs Sourds. Elles sont constituées à partir d'émissions réalisées directement en LS, puis sous-titrées. Ce sont des données d'une très grande qualité à la fois sur la nature de la LS et sur la qualité de l'alignement. Le seul existant à ce jour pour la LSF est le corpus MEDIAPI-SKEL décrit précédemment.

Un autre corpus de ce type a été créé récemment dans le cadre d'une toute première tâche partagée sur la traduction d'une LS vers une langue écrite (LS Suisse Allemande vers l'allemand) dans le cadre

8. https://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/welcome.html
9. https://www.project-easier.eu/
10. https://www.project-easier.eu/wp-content/uploads/sites/67/2021/08/EASIER-D6.1-Overview-of-Datasets-for-the-Sign-Languages-of-Europe.pdf

de la conférence Machine Translation (WMT) [11].

En ce qui concerne le niveau lexical, un des objectifs dans le cadre du projet Easier est constituer une ressource multilingue de type Wordnets pour plusieurs LS. La version actuelle s'est centrée sur les LS grecque et allemande (Bigeard *et al.*, 2022), mais intégrera d'autres LS dont la LSF d'ici la fin du projet.

A l'heure actuelle, une majorité des études, en particulier dans le domaine de la reconnaissance automatique, se concentrent sur les signes lexicaux. Cependant, comme évoqué en section 1, les LS ne sont pas juste une séquence d'unités lexicales discrètes pouvant être répertoriées dans un dictionnaire, mais plutôt des séquences de constructions spatio-temporelles qui combinent des signaux discrets et continus, des composantes manuelles et non manuelles et qui permettent une grande liberté dans la production à la volée d'unités de sens. Les études s'intéressant à ces aspects sont encore très rares.

Pour la LSF, on peut citer les travaux de V. Belissen 2020, dans lesquels l'approche proposée s'est centré sur la détection des unités non lexicales. Pour la partie de son travail portant que la représentation du signeur, au vu du peu de ressources disponibles pour la LSF, il a utilisé un système pré-existantpermettant d'identifier les configurations de mains pour la LS Allemande qu'il a réentrainé pour identifier les configurations de mains sur le corpus DICTA-SIGN (figure 2). Comme son travail n'était pas centré sur le lexique, il est très probablement transférable à d'autres LS.

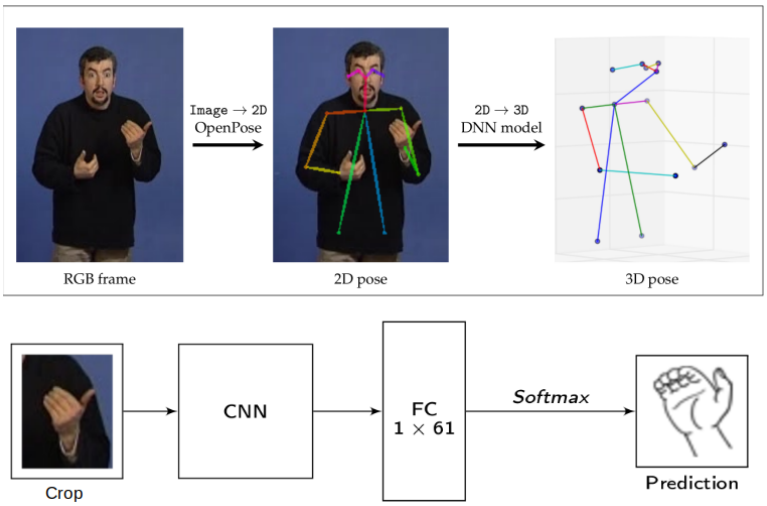


FIGURE 2 – Représentation du signeur en LSF dans la thèse de V. Belissen

En parallèle aux études qui nécessitent des techniques par apprentissage, d'autres approches sont exporées. Par exemple le projet Rosetta [12], terminé récemment, visait à étudier des solutions d'accessibilité pour les contenus audiovisuels. L'un des objectifs consistait à concevoir un système de traduction automatique du texte vers la LSF, affichée via l'animation d'un signeur virtuel. Les trois principales contributions ont été la constitution de ROSETTA-LSF, un corpus aligné de texte et de LS enregistré à l'aide d'un système de capture de mouvement (Bertin-Lemée *et al.*, 2022a), un système de traduction du texte vers une représentation intermédiaire avec une approche à base d'exemples (Bertin-Lemée *et al.*, 2022b), et un système permettant de générer des animations d'avatar à partir de cette représentation et de blocs d'animations préenregistrées (Dauriac *et al.*, 2022). Ce travail a abouti à une preuve de concept fonctionnelle (figure 3) mais limitée par la taille du corpus. Les perspectives

11. https://www.wmt-slt.com/
12. https://rosettaccess.fr/index.php/

de ce travail sont donc liées à la possibilité de développer un corpus aligné de très grande taille ainsi que des outils permettant en particulier de faciliter l'alignement entre la LSF et le français.

FIGURE 3 – Prototype de traduction français vers LSF du projet Rosetta

Un autre aspect important est la possibilité de combiner plusieurs corpus au sein d'un même projet. Par exemple, le projet en cours Serveur Gestuel vise à créer l'équivalent d'un serveur vocal en LSF, intégrant des technologies de reconnaissance, de génération et de dialogue. Pour la partie reconnaissance automatique, les études des laboratoires partenaires du projet sont basées actuellement sur les corpus DICTA-SIGN et MEDIAPI-SKEL. Pour la partie modélisation linguistique qui permet de piloter l'animation de l'avatar, les corpus 40 BREVES et MOCAP1 sont utilisés pour des analyses linguistiques (Martinod *et al.*, 2022).

Ainsi, la possibilité d'utiliser ou produire des systèmes transférables à plusieurs LS et des corpus de différentes LS ou de nature variée permet de dépasser un peu les limites dûes au manque de données.

4 Conclusion

Ces dernières années ont été marquées par des évolutions sur la nature des corpus, la manière de les exploiter ou de les combiner et sur le type de traitement automatique qu'il est possible de mettre en œuvre grâce à eux. En ce qui concerne la LSF, les ressources restent encore très limitées et nécessitent d'être développées et outillée.

Les outils développés en TALS restent à l'heure actuelle des prototypes de recherche et leurs capacités doivent encore être étendues avant de pouvoir procéder à de véritables évaluations. Mais nous pouvons d'ores et déjà souligner que l'évaluation doit elle-aussi être adaptée aux spécificités des LS (multilinéarité, utilisation de l'espace et iconicité), et est un axe de recherche en soi qu'il est nécessaire de développer.

Nom	Description	Taille	Locuteurs	VF	Annot.	Dépôt
LS-COLIN	monologue, narration, récit, explication	1,5h	13	partielle	partielle	Cocoon
CREAGEST	dialogue et acquisition	> 156h *	> 82 *	partielle	partielle	partiel, Ortolang
DEGELS1	dialogue, comparable LSF et gestualité	LSF : 35', gest. : 39'	LSF : 5, gest. : 4	partielle	partielle	Ortolang
40 BREVES	monologue, traduction de brèves journalistiques	1h	3	oui	oui	Ortolang
MOCAP1	monologue, description de photos	2h	8	non	partielle	partiel, Ortolang
DICTA-SIGN-LSF-V2	dialogue, plusieurs tâches avec plusou moins d'élicitation	8h	16	oui	partielle	Ortolang
MEDIAPI-SKEL	très varié, issu d'un média bilingue	27h	>100	oui	non	Ortolang
ROSETTA-LSF	monologue, traduction de phrases de type journalistique	3h	1	oui	oui	Ortolang
CORPUS CATTEAU	poésie en LSF, interprétation en FR et entretiens	*	*	oui	oui	Ortolang
SIGNES EN FAMILLE	Echanges spontanés durant le repas familial	2h à 4h30 par famille	10 familles	non	oui	Ortolang
CLM-MOCAP	monologue et dialogue, capture de mouvement	*	10	non	non	Ortolang
LG-IDF	récit, lexique	1h30	1	non	non	Ortolang

TABLE 1 – Corpus de LSF sur Ortolang et Cocoon. * : non renseigné

Références

BELISSEN V. (2020). *From Sign Recognition to Automatic Sign Language Understanding : Addressing the Non-Conventionalized Units*. Thèse de doctorat. ED STIC, Université Paris-Saclay 2020.

BENCHIHEUB M.-E.-F., BERRET B. & BRAFFORT A. (2016). Collecting and analysing a motion-capture corpus of French Sign Language. In *Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign Languages : Corpus Mining*, p. 7–12, Portorož, Slovenia : ELRA.

BERTIN-LEMÉE E., BRAFFORT A., CHALLANT C., DANET C., DAURIAC B., FILHOL M., MARTINOD E. & SEGOUAT J. (2022a). Rosetta-LSF : an Aligned Corpus of French Sign Language and French for Text-to-Sign Translation. In *Proceedings of the 13th International Conference on Language Resources and Evaluation*, Marseille, France : ELRA.

BERTIN-LEMÉE E., BRAFFORT A., CHALLANT C., DANET C. & FILHOL M. (2022b). Example-Based Machine Translation from Text to a Hierarchical Representation of Sign Language. DOI : 10.48550/ARXIV.2205.03314.

BIGAND F. (2021). *Extracting human characteristics from motion using machine learning : the case of identity in Sign Language*. Thèse de doctorat. ED STIC, Université Paris-Saclay 2021.

BIGEARD S., SCHULDER M., KOPF M., HANKE T., VASILAKI K., VACALOPOULOU A., GOULAS T., DIMOU A.-L., FOTINEA S.-E. & EFTHIMIOU E. (2022). Introducing sign languages to a multilingual wordnet : Bootstrapping corpora and lexical resources of Greek Sign Language and German Sign Language. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages : Multilingual Sign Language Resources*, Marseille, France : ELRA.

BULL H., BRAFFORT A. & GOUIFFÈS M. (2020). MEDIAPI-SKEL - a 2D-skeleton video database of French Sign Language with aligned French subtitles. In *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France : ELRA.

COLLOMB A., BRAFFORT A. & KAHANE S. (2018). L'anatomie du proforme en langue des signes française : Quand il sert à introduire des entités dans le discours. *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence (TIPA)*, (34). DOI : 10.4000/tipa.2164.

DAURIAC B., BRAFFORT A. & BERTIN-LEMÉE E. (2022). Example-based Multilinear Sign Language Generation from a Hierarchical Representation. In *Proceedings of the LREC2022 7th International Workshop on Sign Language Translation and Avatar Technology : The Junction of the Visual and the Textual*, Marseille, France : ELRA.

GARCIA B., L'HUILLIER M.-T. & SALLANDRE M.-A. (2013). Creagest : enjeux linguistiques, patrimoniaux et socio-éducatifs d'un grand corpus de langue des signes française. *La nouvelle revue de l'adaptation et de la scolarisation*, (64). DOI : 10.3917/nras.064.0081.

MARTINOD E., DANET C. & FILHOL M. (2022). Two new AZee production rules refining multiplicity in French Sign Language. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages : Multilingual Sign Language Resources*, Marseille, France : ELRA.

SALLANDRE M.-A., BALVET A., BESNARD G. & GARCIA B. (2019). Étude exploratoire de la fréquence des catégories linguistiques dans quatre genres discursifs en LSF. *Revue de linguistique et de didactique des languese (LIDIL)*, (60). DOI : 10.4000/lidil.7136.

Pull your treebank up by its own bootstraps

Ziqian Peng[1] Kim Gerdes[1] Kirian Guiller[2]

(1) Lisn (CNRS), Université Paris-Saclay, France
(2) Modyco (CNRS), Université Paris-Nanterre, France

ziqian.peng@universite-paris-saclay.fr
kim.gerdes@universite-paris-saclay.fr
kguiller@parisnanterre.fr

RÉSUMÉ

Remontez les bretelles à votre treebank

Nous analysons la performance de récents analyseurs syntaxiques neuronaux dans la tâche d'amorçage d'un treebank, c'est-à-dire l'entraînement et l'analyse itérative afin d'améliorer la vitesse et la qualité de l'analyse syntaxique humaine. En effectuant une recherche extensive et heuristiquement guidée dans la vaste grille d'options (analyseur syntaxique, plongement, configuration, époques, taille du batch, taille de l'ensemble d'entraînement, schéma d'annotation, langue, méthode d'évaluation...), nous déterminons les configurations d'analyseurs syntaxiques les plus performantes : UDify et Trankit se partagent le podium en fonction de la taille de l'ensemble d'entraînement. Nous montrons également comment ces résultats sont intégrés dans l'outil d'annotation ArboratorGrew, et nous proposons quelques mesures préliminaires qui permettent de prédire la qualité de l'analyse syntaxique pour une nouvelle langue.

ABSTRACT

We analyze the performance of recent neural syntactic parsers in the task of bootstrapping a treebank, i.e. training and analyzing iteratively in order to enhance speed and quality of the human syntactic analysis. By conducting an extensive and heuristically guided search in the vast grid of options (parser, embedding, configuration, epochs, batch size, size of training set, annotation scheme, language, evaluation method…), we determine the best performing parser configurations: UDify and Trankit share the podium depending on the size of the training set. We also show how these results are integrated into the annotation tool ArboratorGrew, and we propose some preliminary measures that allow predicting the quality of the parse for a new language.

MOTS-CLÉS : treebanks, annotation, analyseurs syntaxiques, réseaux neuronaux, amorçage, langues sous-ressourcées.

KEYWORDS: treebanks, annotation, syntactic parsers, neural networks, bootstrapping, under-resourced languages.

1 Introduction

Treebanks are steadily gaining importance as a tool for conducting research in syntax but their development is resource hungry in researcher's working hours as well as in the development and usage of recent neural network based tools. This is one of the reasons why the set of languages in Universal Dependencies (UD) is heavily biased towards well-resourced languages although an increasing number of, albeit often small, treebanks are developed for lower-resource languages (see for example the TowerParse project, Glavaš and Vulić 2021). This fact limits the scope of typological data-based studies on treebanks. In the context of the ANR project Autogramm (2022-2025), we develop a set of new treebanks for low-resource languages in the SUD annotation scheme (Gerdes et al. 2018, 2019, 2021). It is easier to annotate in SUD when no pre-established grammar exists that would provide a distinction between content and function words, which is commonly the case for less-resourced languages. Furthermore, SUD has been shown to be cognitively more relevant (Yan and Liu 2019), and parser performance improves on function-word-as-head annotation schemes (Rehbein et al. 2017)

We want to provide the usually less computer-inclined field linguists with state-of-the-art tools to develop high-quality treebanks and thus fill some gaps in treebank-based typological studies. More concretely, we want to answer the common questions of any syntactician wanting to start a new treebank: How many sentences do I have to annotate before it makes sense to train a first model? What parsing quality can I expect? How often should I retrain and reparse? What parser, embedding, and configuration should I use? How long does it take on my GPU? Can we make educated guesses on these questions based on raw or POS-tagged text?

2 Analysis and results

Although syntactic parsers are less relevant than they used to be for many NLP downstream tasks, these tools are still under very active development, in particular in a linguistic or low-resource perspective, and finding the parser best fitting for a given task is a quickly moving target. We chose 5 recent parsers: UDify (Kondratyuk and Straka 2019), Hopsparser (Grobol and Crabbé 2021), Trankit (Nguyen et al. 2021), Stanza (Qi et al 2020), BertForDeprel (Guiller 2020) and we tested their performance on 5 typologically diverse languages, English (en), French (fr), Chinese (zh), Japanese (ja) and Arabic (ar). We tested the parsers for six training sizes with the number of sentences [10, 30, 50, 100, 300, 500] during 100 epochs with 10-fold cross-validation, which gives us 5x5x6x10=1500 models to train and to evaluate. These numbers of sentences seem to us to be a reasonable grid for bootstrapping during the annotation process of a treebank for a new language.

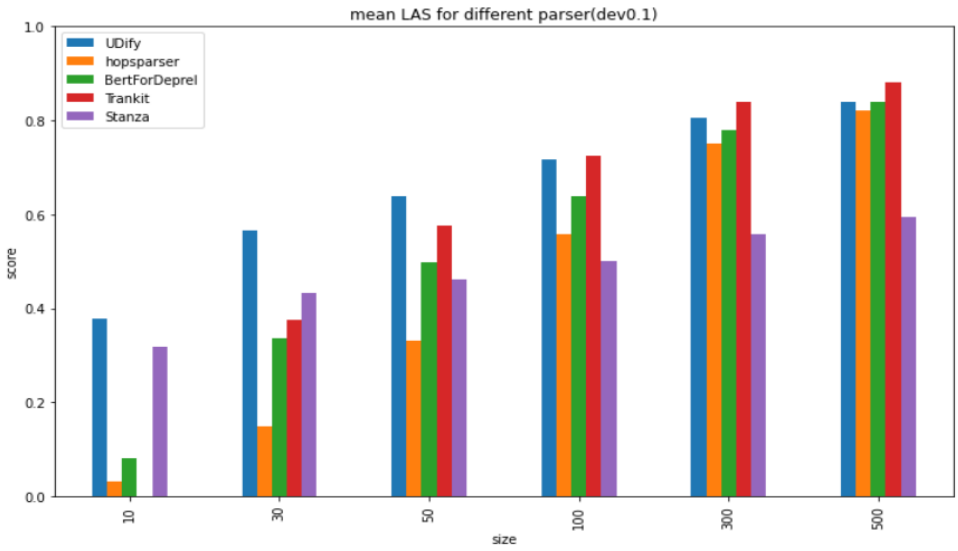


Figure 1: average parser performance for the 5 test languages: Training size vs. Labeled Attachment Score (LAS)

To do this, we randomly selected from SUD v2.10[1] 1500 sentences for each language, with 500 for training and 1000 sentences as test files, to be parsed by each trained parser and evaluated with the official UD evaluation script, so as to make the evaluation scores comparable.

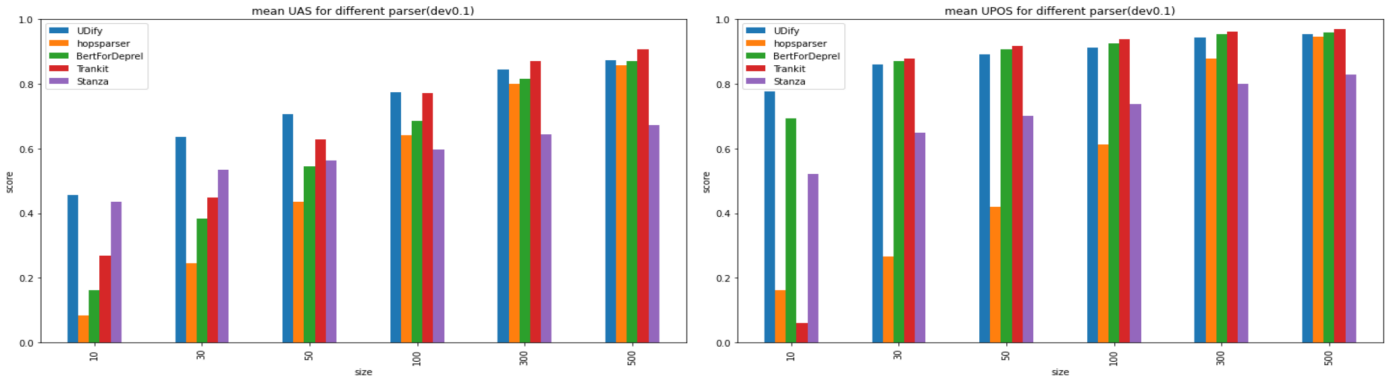


Figure 2: average parser performance for the 5 test languages: Training size vs. Unlabeled Attachment Score (UAS) at left and Training size vs. Universal POS tag (UPOS) at right

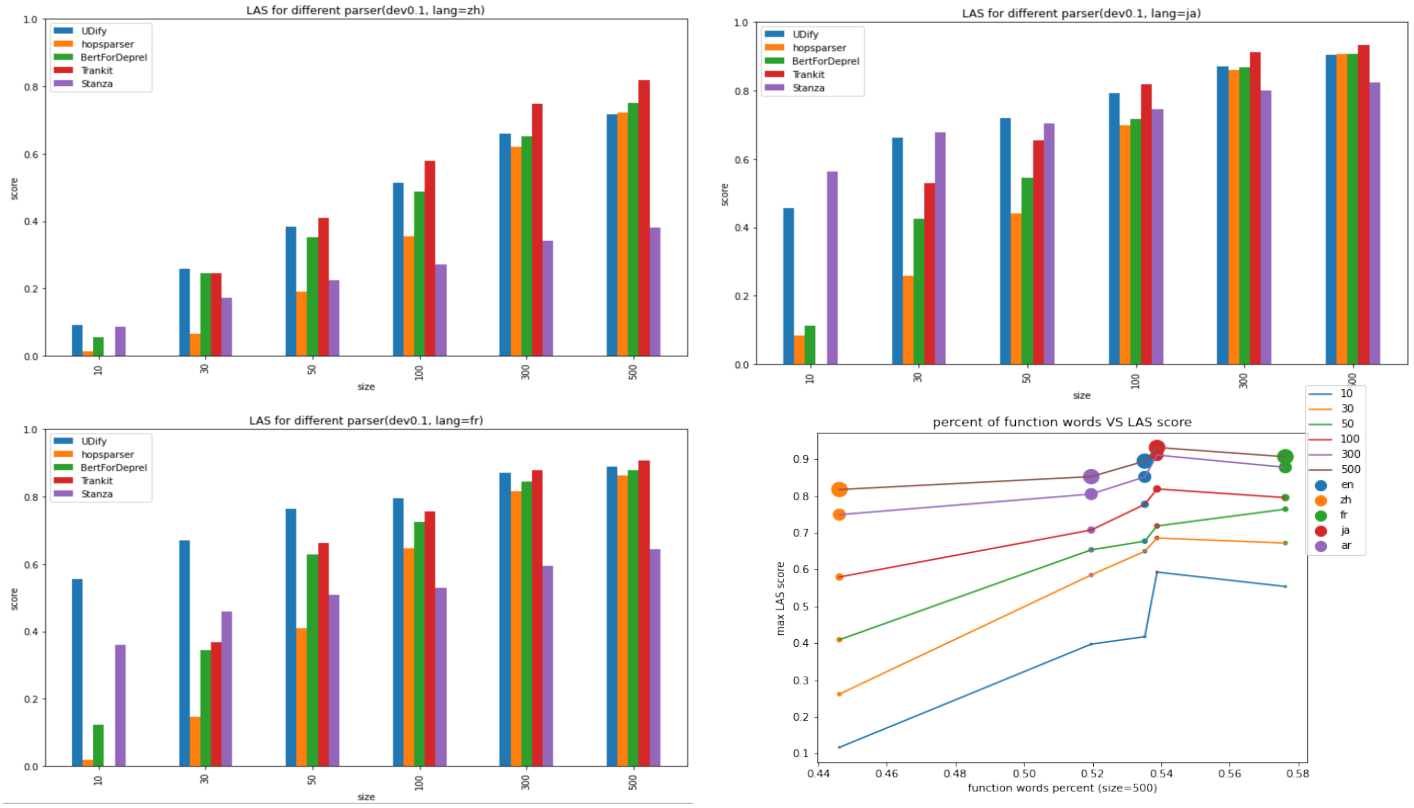


Figure 3: average parser performance for the 5 test languages: Training size vs. Labeled Attachment Score (LAS) for Chinese (first row at left, the worst performance as LAS), Japanese (first row at right, the best performance as LAS), French (second row at left) and the percent of functional words vs max LAS F1 score per language per size.

Before diving into the details, let us have a look at the average LAS (Labeled Attachment Score) results across our 5 languages: Figure 1 shows that UDify is clearly ahead when trained on a small training set. Starting with 100 sentences, Trankit takes the lead. These results are corroborated in the more detailed results below.

1 https://surfacesyntacticud.github.io/data/ In order to gather enough data, we had to merge several treebanks of each language.

2.1 Detailed parsing results for the 5 base languages

Just as for LAS, UDify starts the race of Unlabeled Attachment Scores (UAS) trained on very few sentences, but it is overtaken by Trankit only at 300 sentences. For precision during POS tagging, Trankit already takes the lead with 30 sentences to train on. When distinguishing the results by language, we first note that LAS takes more training data on Chinese than on other languages to reach comparable scores. Japanese and French, on the contrary, have above-average performance in LAS.

In the last graph of Figure 3 on the right, we ordered the languages by their percentage of function words, from 45% for Chinese (zh) to 57% for French (fr). As expected, we observe a general tendency of faster learning in languages with more function words, but the results for French are less good than for Japanese although it has more function words. Note that differences between the languages get less prominent the more training sentences we have.

2.2 Detailed parsing results for all available SUD treebanks

These first tests were based on the above-mentioned 5 languages. Based on these results, we repeated limited tests on the 69 languages of SUD 2.10 where 1500 sentences are available: We only tested on the two best-performing parsers Trankit and UDify, and we did not perform cross-validation.

When looking at Trankit's and UDify's performance per simple dependency relation (grouping subrelations, such as comp:ob under comp, see confusion matrices in the Annex), we see that the worst scores appear for the rare relations such as *orphan*, *reparandum,* and *list*) with a precision of 34%, 57% and 60% respectively (36%, 26%, and 72% respectively for UDify). Trankit's highest confusion rate is found for *udep* vs *orphan* with 18%. For UDify it is *reparandum* vs. *root* that causes a 34% confusion, pointing to a different tree spanning algorithm to create the trees.

For LAS, UAS, and POS tagging, we measure the average of both parsers. The results show very high discrepancies between the languages, ranging from 90% LAS for Greek to 22% for Coptic. The ancient languages are characterized by the fact that they are not easy to parse. It is not readily possible to determine the cause of these results, as it may be the genre of the texts, the languages themselves, or the incoherent annotation that the parser cannot pick up. It is noticeable that no language group stands out as being particularly easy to parse.

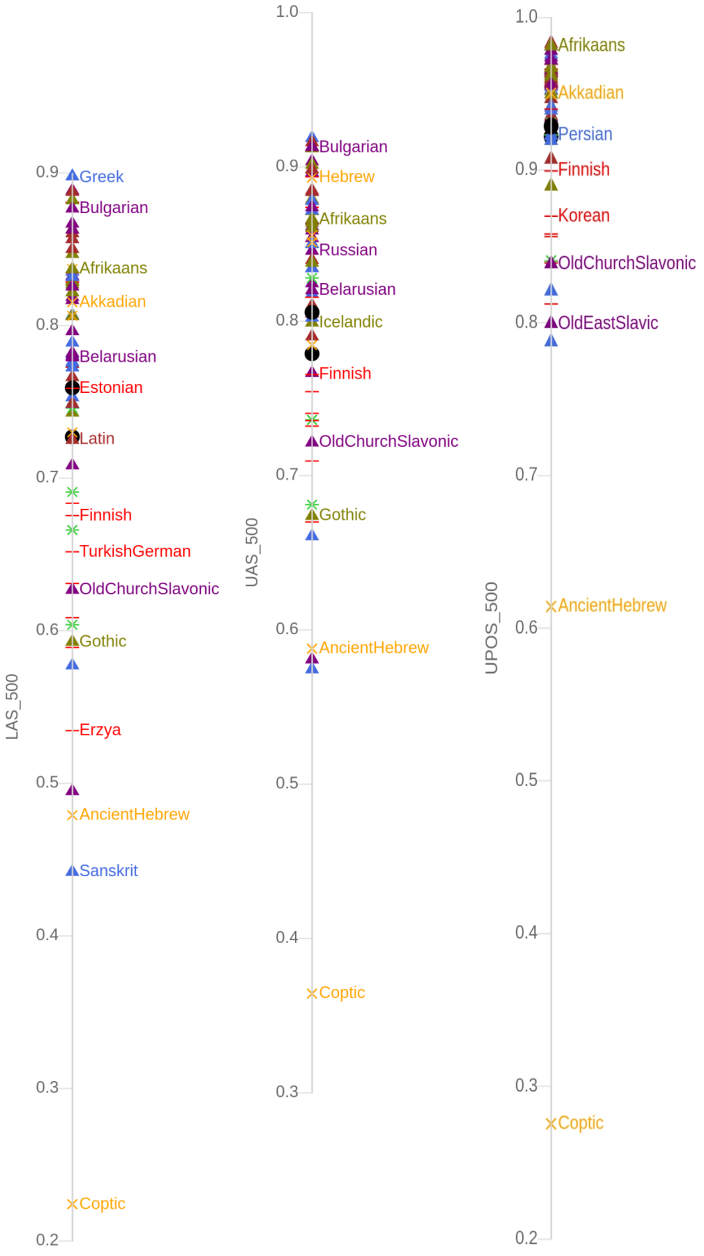


Figure 4: Average performance of Trankit and UDify measured by LAS (left), UAS (middle), UPOS (right) on dataset of size 500 for each of the 69 languages

Unsurprisingly, the LAS performance of the parsers is highly correlated with the POS accuracy. Put differently, as soon as we know the performance on POS tagging, we can predict reasonably well the LAS performance of the syntactic parser. Our data suggests that for 500 sentences in the language L, the LAS score can be computed by LAS(L)=1.55*POS-0.68.

The parser performance measured on the 69 languages confirms what we have observed before on our 5 test languages: Trankit needs at least 100 training sentences to catch up to Udifiy, but then delivers better parser results. See Figure A in Annex I for a graphical representation of these measures.

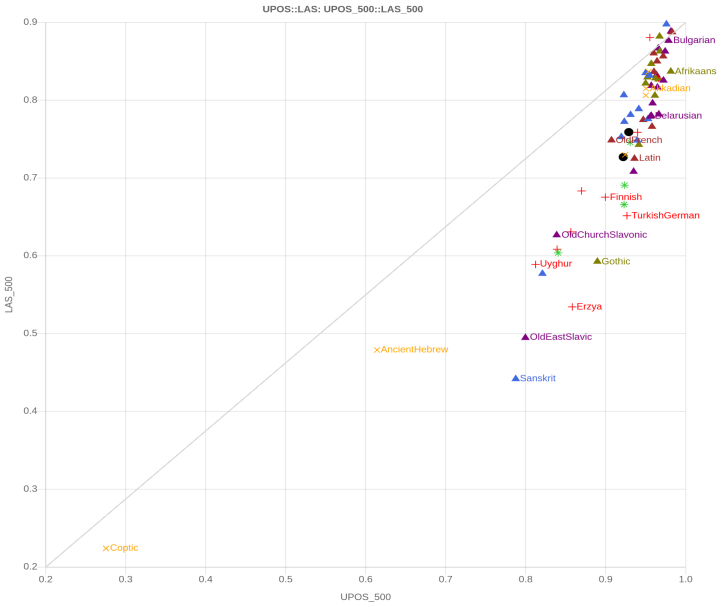


Figure 5: average performance of Trankit and UDIfy per language on dataset of size 500: Universal POS tag (UPOS) vs labeled Attachment Score (LAS)

3. Predicting parser performance

The observed significant differences in parser performance make it hard to give general predictions on the parser performance during treebank bootstrapping. Are there other measures that can be performed on raw or POS-tagged texts that can help us make better predictions? In this section we will show how the type/token ratio and the percentage of function words influences the parser performance, which allows us to make predictions of parser performance based on these measures. These findings are implemented as heuristics for automatically tweaking the parser parameters to optimize parser performance in the ArboratorGrew annotation tool (Guibon et al. 2020).

3.1 Can the type/token ratio predict parser performance?

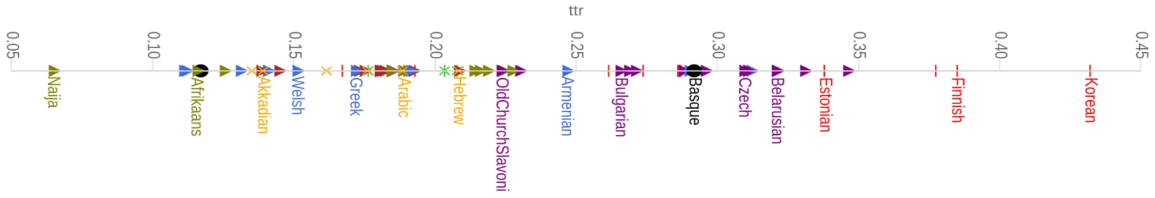


Figure 6: Type/Token ratio of the 69 languages, evaluated on the 1500 selected sentences.

The type/token ratio (TTR) is a measure of lexical richness of a text. As the TTR decreases for all languages with the size of the text, we need to measure it on texts of approximately the same length in order to make it comparable. The plot of Figure 6 shows large differences between the languages, the lead being taken by agglutinating languages followed by Slavic languages, and Korean being the "richest" language with a TTR of 43%. The large TTR difference between structurally similar languages such as Korean and Japanese (at only 19% not shown above) can be explained by different word segmentation rules underlying the treebanks: Japanese is separating the verbal and nominal suffixes, resulting in many equal functional tokens, and Korean considers the suffixes as

part of the word, resulting in many unique tokens.

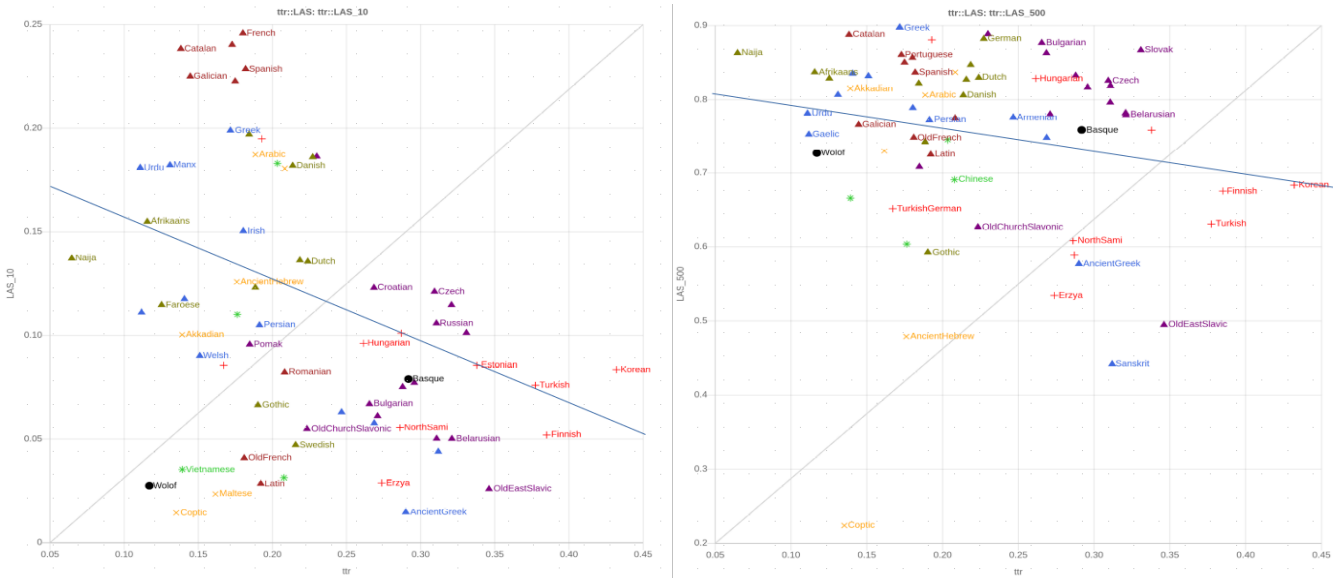


Figure 7: Type/Token Ration (TTR) vs LAS on dataset of size 10 (left) and that of size 500 (right), with blue lines illustrating correlation between TTR and LAS (cf. Annex VI)

As expected, we observe a negative correlation between TTR and LAS: The richer the language the harder it is to parse. Also, the Spearman correlation coefficient decreases between scores for training on 10 and on 500 sentences, respectively -0.33 and -0.17, indicating that the measure becomes less relevant with larger training sets.

3.2 Can POS tags predict parser performance?

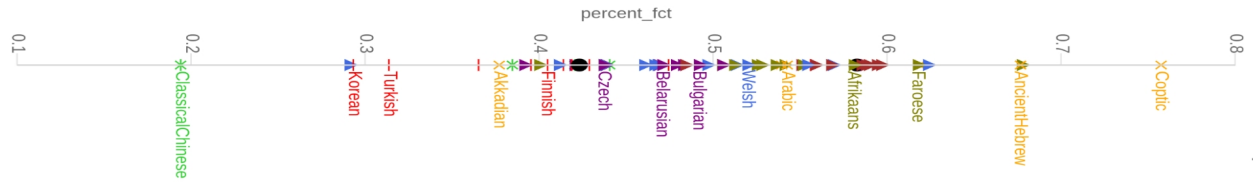


Figure 8: overview of the percentage of function words across languages, evaluated on the 1500 selected sentences.

We have shown above that POS tagging performance is a very good indicator of LAS performance. But can the distribution of POS themselves be a predictor? Our hypothesis is that the lexical vs. function word distinction allows us to make predictions: The more function words, the easier. Taking nouns, verbs, adjectives, and adverbs as lexical categories, we first observe a distribution ranging from 20% function words in classical Chinese to 76% in Coptic. Plotting these measures against the LAS score, we observe the expected positive correlation. The two languages with the highest percentage of function words, Coptic and Ancient Hebrew, are outliers of the general tendency. The Spearman correlation coefficient is 0.4453 for 500 sentences.

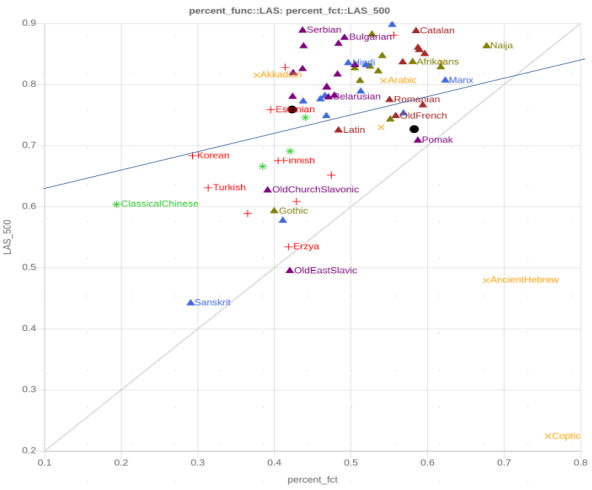


Figure 9: Percentage of function words of the 69 languages computed with selected sentences VS LAS F1 score on dataset of size 500

3.3 Parser performance and language structure

Two other interesting results are the measures of language directionality and tree height: When measuring tree directionality (the average dependency length, counting leftward relations negatively), we observe that this measure has little influence on the parser results (Spearman correlation 0.3230 with p-value 7.6% > 5% to accept the null hypothesis that the observed correlation is due to chance). Inversely, the tree height has a very profound influence on the parser: the higher the tree, the better the score (Spearman 0.5126). This latter correlation may be another explanation for the better parser performance of SUD vs UD: SUD's function word centric approach simply results in higher trees. The integration of the parser results into the typometrics platform https://typometrics.elizia.net allows for further study of the correlation between various treebank measures and treebank performance.

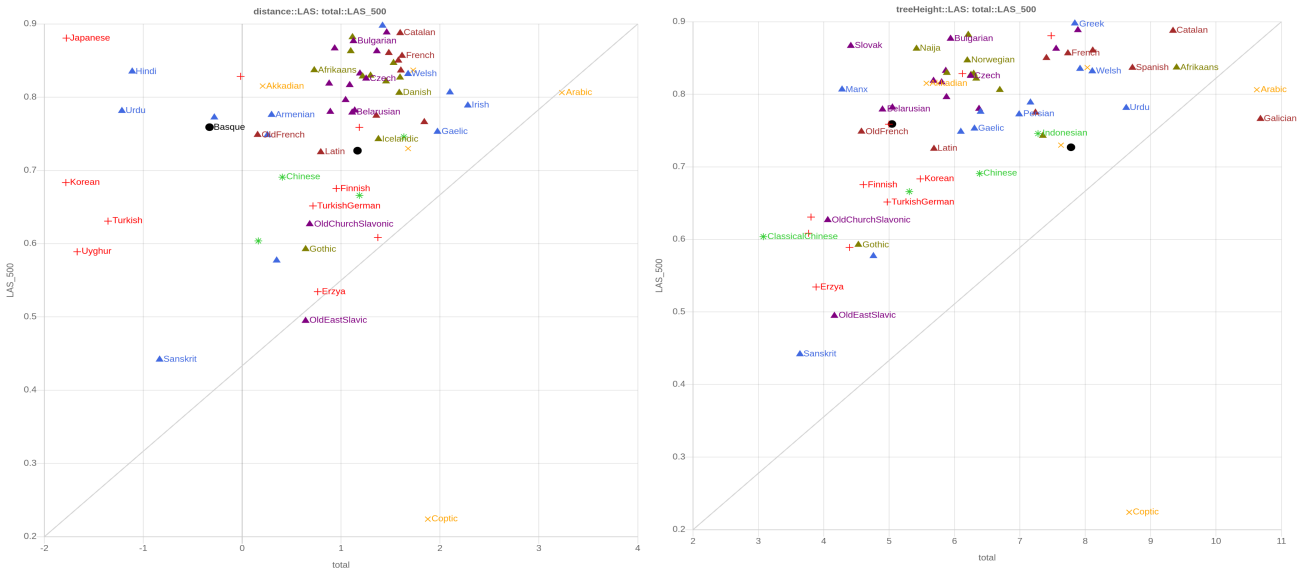


Figure 10: Distance computed on SUD2.8 VS LAS F1 score on dataset of size 500 (left) and treeheight computed on SUD2.8 VS LAS F1 score on dataset of size 500 (right)

4. Implementation in ArboratorGrew

ArboratorGrew is a new treebank annotation tool that integrates Grew's graph search and rewrite features into Arborator's collaborative online annotation platform. The new train-and-parse option makes it possible to use any of the five parsers to train a model on some samples, and obtain the parse results on other samples. The parser operates on a separate server equipped with a high-performance graphics card.[2] The interface proposes simple options and makes predictions on the required time to train and parse based on a logical regression, see the Annex for a screenshot and for the time regression lines.

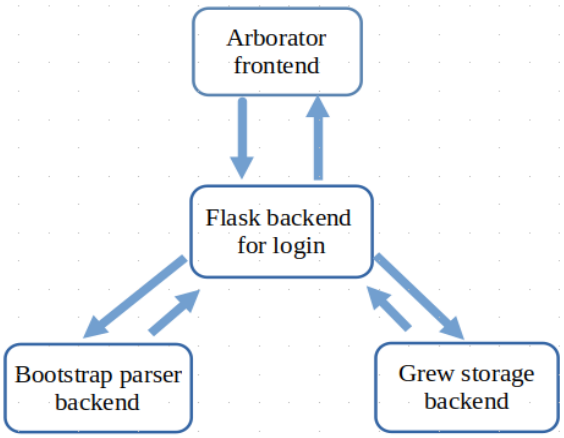


Figure 11: architecture of frontend and backend.

2 The bootstrapping backend relies on a single Nvidia RTX A6000 card with 48Gb of RAM.

5. Conclusion

The influence of annotating pre-analyzed text has been discussed in Fort & Sagot 2010, and we should be aware that the syntactician is less likely to detect new peculiar constructions when presented with a reasonably well pre-annotated text. On the other hand, the lower diversity of the pre-analyzed treebank annotation naturally results in better parser performance.

The rather quick increase in LAS with the size of the training set suggests a very early and regular bootstrapping approach, possibly best starting with 30 sentences. When exactly this makes sense heavily depends on the language, and, of course, on the time measures on parse corrections compared to an annotation from scratch. With such a measure, which remains to be done on a variety of annotators' profiles, it would be possible to answer for example whether a pre-annotation with only 50% LAS is still useful or not.

We also see the need to improve the ArboratorGrew tool: The "diff" mode showing the difference between two trees should show the certainty of proposed relations, so as to allow the annotator to see directly the problematic relations that require scrutiny. Also, a single manual correction should optionally trigger a recomputation of the minimum spanning tree so that the most likely structure, given the new relation, can be proposed directly without further manual intervention. This would significantly reduce the correction time spent on faulty parse results.

Acknowledgements

We would like to thank our anonymous reviewers for their interesting remarks and questions. Loïc Grobol helped with the implementation of the Hops parser and pointed us to Trankit. Laurent Pointal helped us to develop a secure parser backend, and the intensive parser training was done on both the Lisn and the Lab-ia clusters at the University Paris-Saclay.

References

Fort, Karën, and Benoît Sagot. "Influence of pre-annotation on POS-tagged corpus development." In The fourth ACL linguistic annotation workshop, pp. 56-63. 2010.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, Guy Perrier. Starting a new treebank? Go SUD! Theoretical and practical benefits of the Surface-Syntactic distributional approach in DepLing 2021.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, Guy Perrier. Improving Surface-syntactic Universal Dependencies (SUD): surface-syntactic relations and deep syntactic features in TLT 2019.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, Guy Perrier. SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD in UDW 2018.

Goran Glavaš and Ivan Vulić, “Climbing the tower of treebanks: Improving low-resource dependency parsing via hierarchical source selection,” in Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 4878–4888.

Loıc Grobol and Beno t ıCrabbé, “Analyse en dépendances du français avec des plongements contextualisés,” in TALN/RECITAL 2021, 2021.

Guibon, Gaël, Marine Courtin, Kim Gerdes, and Bruno Guillaume. "When collaborative treebank curation meets graph grammars." In LREC 2020 -- 12th Language Resources and Evaluation Conference. 2020.

Kirian Guiller. "Analyse syntaxique automatique du pidgin-créole du Nigeria à l’aide d’un transformer (BERT): Méthodes et Résultats." Mémoire de Master, Sorbonne Nouvelle (2020).

Dan Kondratyuk and Milan Straka, “75 languages, 1 model: Parsing universal dependencies universally,” arXiv preprint arXiv:1904.02099, 2019.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen, “Trankit: A light-weight transformer-based toolkit for multilingual natural language processing,” arXiv preprint arXiv:2101.03289, 2021.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning, “Stanza: A python natural language processing toolkit for many human languages,” arXiv preprint arXiv:2003.07082, 2020.

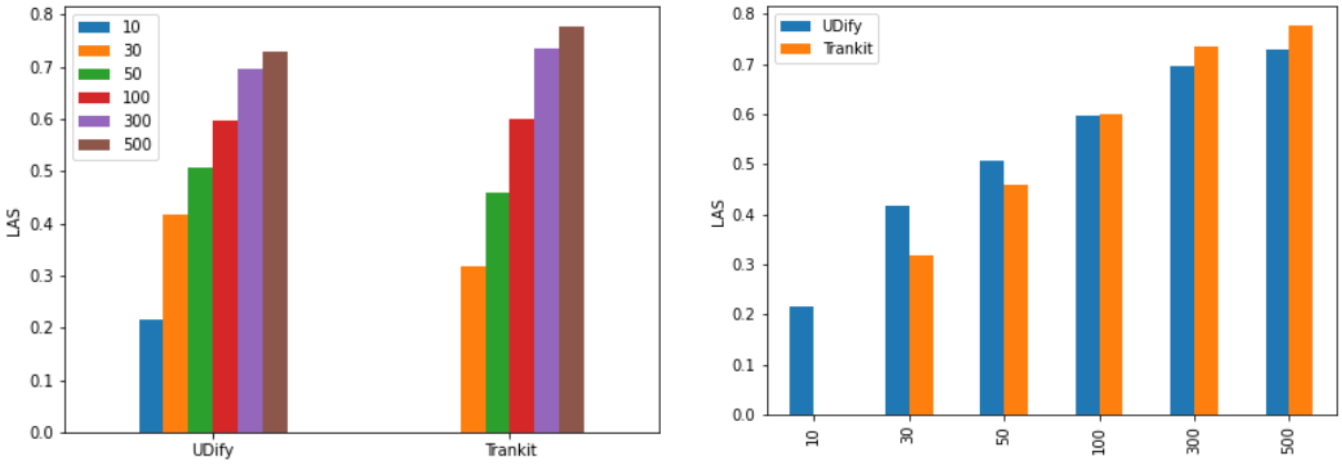
Rehbein, Ines, Julius Steen, Bich-Ngoc Do, and Anette Frank. "Universal Dependencies are hard to parse–or are they?." In Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), pp. 218-228. 2017.

Yan, Jianwei, and Haitao Liu. Which annotation scheme is more expedient to measure syntactic difficulty and cognitive demand?. Presented at Quasy, SyntaxFest 2019.

Annex

I. Statistic of LAS f1 score for Trankit and UDify

Figure A: Performance of UDify and Trankit for the 69 available languages in SUD2.10 : parser vs Labeled Attachment Score (LAS) at left and Training size vs LAS at right



Statistic of LAS f1 score for Trankit (left) and UDify (right) on 69 languages

	10	30	50	100	300	500
count	69.0	69.000000	69.000000	69.000000	69.000000	69.000000
mean	0.0	0.318216	0.457290	0.600810	0.735069	0.776311
std	0.0	0.083658	0.120591	0.131782	0.125436	0.117720
min	0.0	0.154317	0.170476	0.195329	0.219645	0.219452
25%	0.0	0.253727	0.377353	0.528295	0.672990	0.731561
50%	0.0	0.327911	0.468934	0.639530	0.776533	0.806653
75%	0.0	0.381771	0.556436	0.689709	0.820712	0.850298
max	0.0	0.475108	0.653467	0.788150	0.885604	0.914372

	10	30	50	100	300	500
count	69.000000	69.000000	69.000000	69.000000	69.000000	69.000000
mean	0.220727	0.416438	0.508161	0.595730	0.694433	0.727304
std	0.127837	0.153908	0.156169	0.153074	0.140600	0.134007
min	0.028970	0.122755	0.166573	0.210142	0.223761	0.228609
25%	0.115289	0.303778	0.388442	0.508322	0.630197	0.670072
50%	0.202076	0.409790	0.530283	0.636877	0.741292	0.767542
75%	0.309760	0.544047	0.648022	0.714934	0.793280	0.816323
max	0.491621	0.680553	0.761930	0.805829	0.861106	0.881114

Languages with LAS less than 0.5 when the dataset contains 500 sentences: Trankit is more universal than UDify so that only 2 languages got LAS less than 0.5 with dataset of size 500.

```
las_trankit[las_trankit['500'] < 0.5]
```

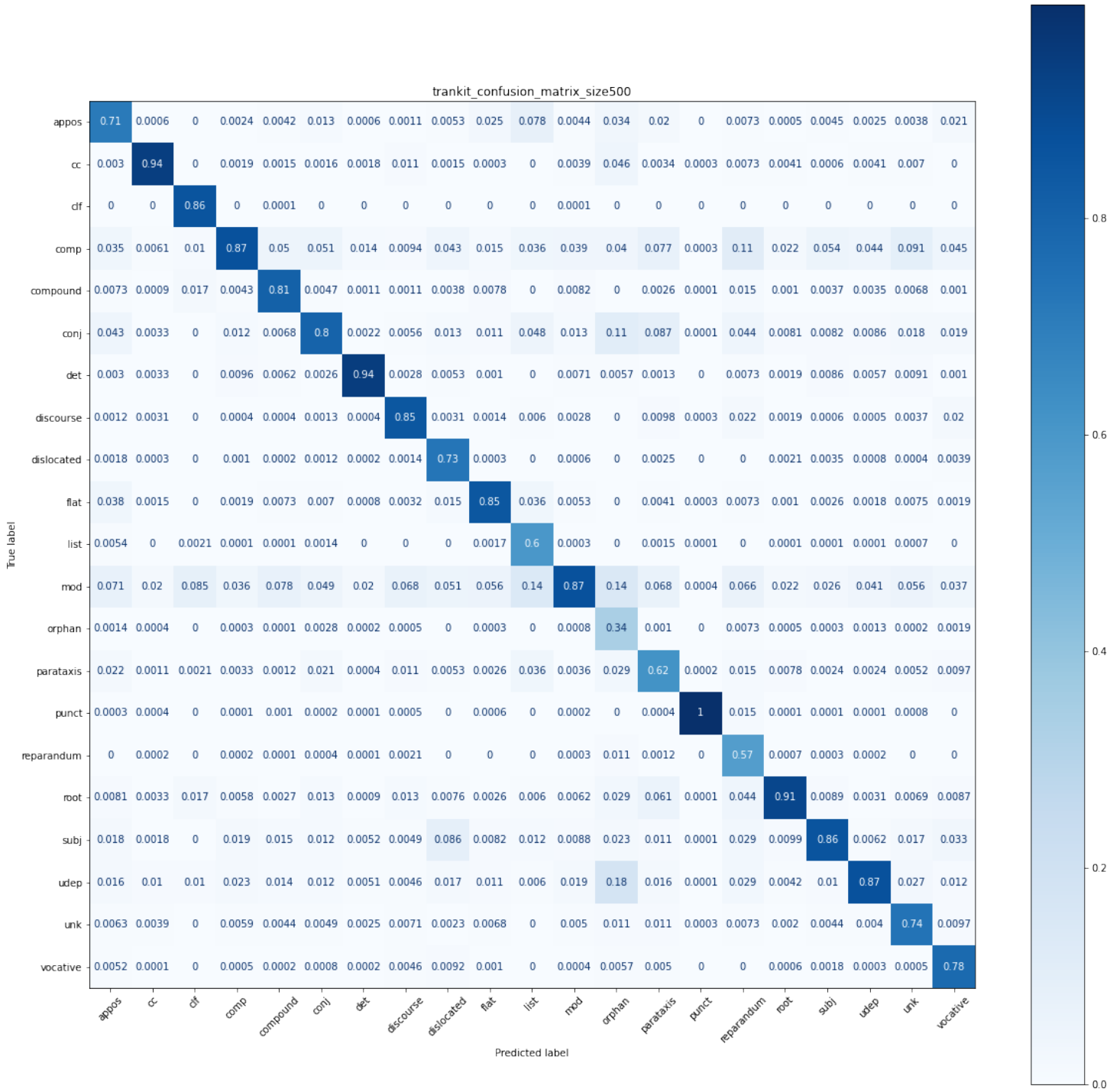
	10	30	50	100	300	500
cop	0.0	0.154317	0.170476	0.195329	0.219645	0.219452
sa	0.0	0.159379	0.200846	0.252045	0.390550	0.486037

```
las_udif[las_udif['500'] < 0.5]
```

	10	30	50	100	300	500
cop	0.028970	0.168629	0.184672	0.210142	0.223761	0.228609
grc	0.029599	0.188796	0.200606	0.304921	0.429826	0.477139
hbo	0.251773	0.295608	0.332957	0.364424	0.393916	0.407937
orv	0.051763	0.166951	0.196107	0.275949	0.391675	0.433659
sa	0.087870	0.144147	0.166573	0.226375	0.329055	0.397743
ug	0.202076	0.218613	0.299190	0.347368	0.432838	0.471903

II. Confusion matrix for Precision

II.1 Trankit:



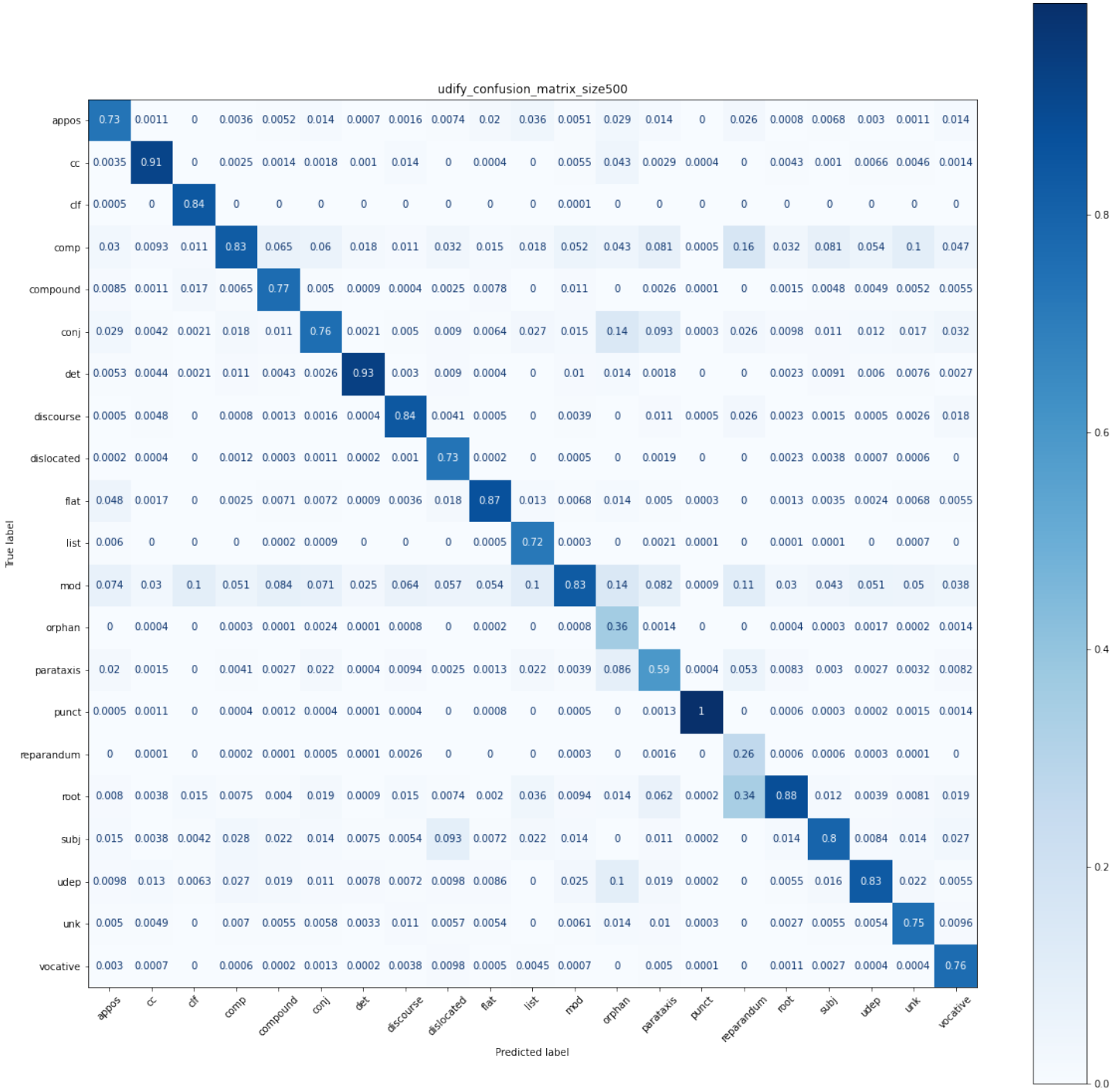
For Trankit std < 0.08 (left), std > 0.15: (right)

	cc	clf	det	punct	udep
std	0.034557	0.07126	0.032405	0.004537	0.05992

	appos	conj	list	parataxis	reparandum
std	0.174365	0.217287	0.166138	0.157149	0.158298

This table shows the standard variation over the different training sizes from 10 to 500 for various relations. E.g. The lowest standard variation is for *cc* which indicates that the analysis does not improve significantly with a bigger training set. On the contrary, rare relations such as appos, conj, list, parataxis, and referendum are still varying a lot and can be expected to improve with a larger training set. Check also the last column of F1 score for Trankit and UDify in Annex III.

II.2 UDify



Amount of different Deprel in the 69 languages:

	appos	cc	clf	comp	compound	conj	det	discourse	dislocated	flat	goeswith	list	mod	orphan	parataxis	punct	reparandum	root	subj	udep	unk	vocative
amount	9152	45089	447	320594	21454	51185	81468	6282	2139	15809	35	393	235847	768	8729	151395	368	69000	89670	123638	13101	1517

We exclude *goeswith* from the confusion matrix since there are only 35 occurrences of this relation.

III. F1 score for Trankit(left) and UDify(right) by size

	10	30	50	100	300	500	500-50
appos	0.0	0.162	0.207	0.390	0.560	0.632	0.425
cc	0.0	0.847	0.872	0.899	0.931	0.938	0.066
clf	0.0	0.611	0.741	0.822	0.878	0.892	0.150
comp	0.0	0.682	0.752	0.816	0.874	0.891	0.139
compound	0.0	0.539	0.612	0.700	0.772	0.790	0.178
conj	0.0	0.228	0.357	0.536	0.736	0.792	0.435
det	0.0	0.818	0.851	0.889	0.919	0.928	0.077
discourse	0.0	0.461	0.534	0.646	0.769	0.805	0.270
dislocated	0.0	0.344	0.364	0.417	0.523	0.556	0.192
flat	0.0	0.375	0.510	0.644	0.774	0.811	0.301
goeswith	0.0	0.000	0.000	0.000	0.000	0.089	0.000
list	0.0	0.019	0.034	0.129	0.311	0.354	0.320
mod	0.0	0.650	0.714	0.782	0.849	0.871	0.157
orphan	0.0	0.009	0.017	0.043	0.072	0.125	0.108
parataxis	0.0	0.181	0.259	0.341	0.478	0.541	0.283
punct	0.0	0.989	0.993	0.996	0.997	0.998	0.005
reparandum	0.0	0.116	0.153	0.198	0.234	0.309	0.156
root	0.0	0.700	0.780	0.841	0.893	0.909	0.129
subj	0.0	0.573	0.678	0.759	0.837	0.862	0.184
udep	0.0	0.737	0.779	0.814	0.856	0.871	0.092
unk	0.0	0.273	0.378	0.496	0.621	0.663	0.286
vocative	0.0	0.191	0.295	0.365	0.521	0.627	0.333

	10	30	50	100	300	500	500-50
appos	0.050	0.177	0.276	0.387	0.536	0.579	0.303
cc	0.643	0.824	0.844	0.873	0.907	0.915	0.071
clf	0.423	0.671	0.780	0.817	0.855	0.869	0.089
comp	0.524	0.667	0.723	0.778	0.838	0.856	0.133
compound	0.198	0.462	0.536	0.617	0.704	0.730	0.194
conj	0.085	0.255	0.379	0.523	0.680	0.733	0.354
det	0.702	0.798	0.833	0.868	0.904	0.913	0.080
discourse	0.290	0.420	0.497	0.580	0.704	0.743	0.245
dislocated	0.074	0.219	0.249	0.389	0.494	0.532	0.283
flat	0.244	0.435	0.549	0.645	0.765	0.793	0.244
goeswith	0.000	0.000	0.000	0.000	0.000	0.000	0.000
list	0.021	0.064	0.052	0.156	0.288	0.519	0.468
mod	0.425	0.588	0.665	0.730	0.803	0.828	0.163
orphan	0.000	0.009	0.020	0.019	0.054	0.060	0.040
parataxis	0.053	0.194	0.251	0.320	0.443	0.494	0.243
punct	0.844	0.981	0.988	0.992	0.995	0.996	0.008
reparandum	0.005	0.040	0.034	0.033	0.050	0.049	0.016
root	0.492	0.686	0.744	0.803	0.862	0.880	0.136
subj	0.305	0.501	0.598	0.679	0.770	0.800	0.203
udep	0.568	0.696	0.733	0.776	0.822	0.839	0.106
unk	0.066	0.236	0.332	0.438	0.559	0.611	0.280
vocative	0.177	0.251	0.322	0.349	0.433	0.497	0.175

Note that we cannot train the Trankit pipeline for our dataset of language ga (Irish) with only 10 sentences. The last column of both tables reporte the improvement of F1 score with the augmentation of data size from 50 to 500. The score for conj, appos and list have been improved more than 30% with both parsers.

IV. List of the 69 languages :

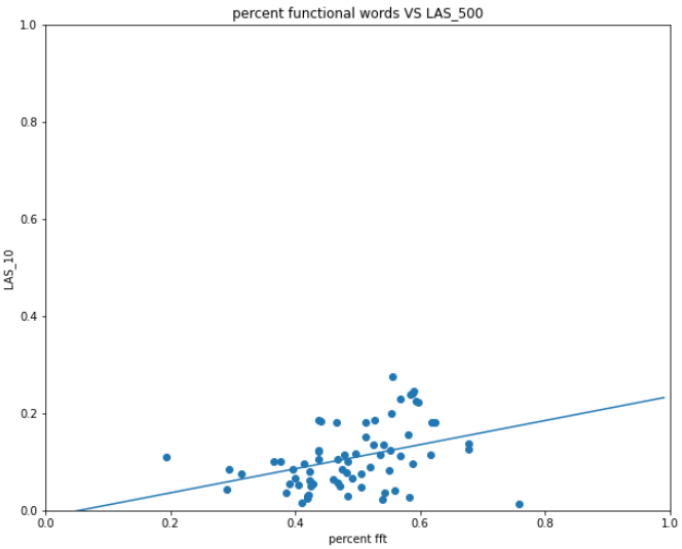
Afrikaans, Akkadian, AncientGreek, AncientHebrew, Arabic, Armenian, Basque, Belarusian, Bulgarian, Catalan, Chinese, ClassicalChinese, Coptic, Croatian, Czech, Danish, Dutch, English, Erzya, Estonian, Faroese, Finnish, French, Gaelic, Galician, German, Gothic, Greek, Hebrew, Hindi, Hungarian, Icelandic, Indonesian, Irish, Italian, Japanese, Korean, Latin, Latvian, Lithuanian, Maltese, Manx, Naija, NorthSami, Norwegian, OldChurchSlavonic, OldEastSlavic, OldFrench, Persian, Polish, Pomak, Portuguese, Romanian, Russian, Sanskrit, Serbian, Slovak, Slovenian, Spanish, Swedish, Turkish, TurkishGerman, Ukrainian, Urdu, Uyghur, Vietnamese, Welsh, WesternArmenian, Wolof

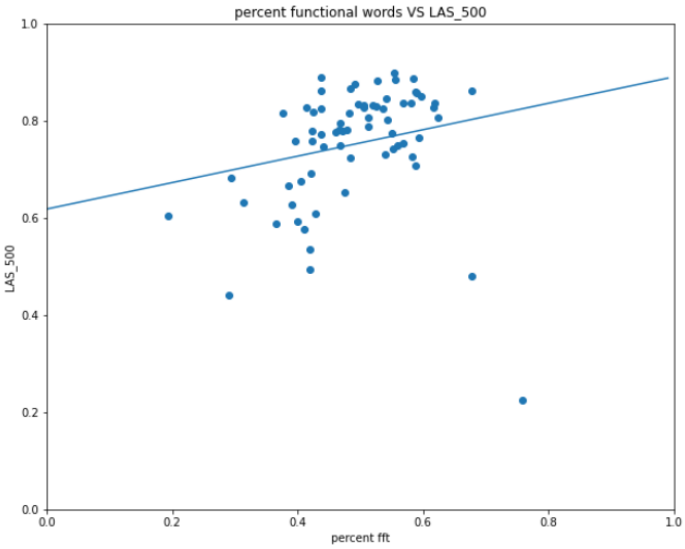
V. Parser configuration:

For all parsers 10% for the dev set after comparison between 10%, 20% and 30%.

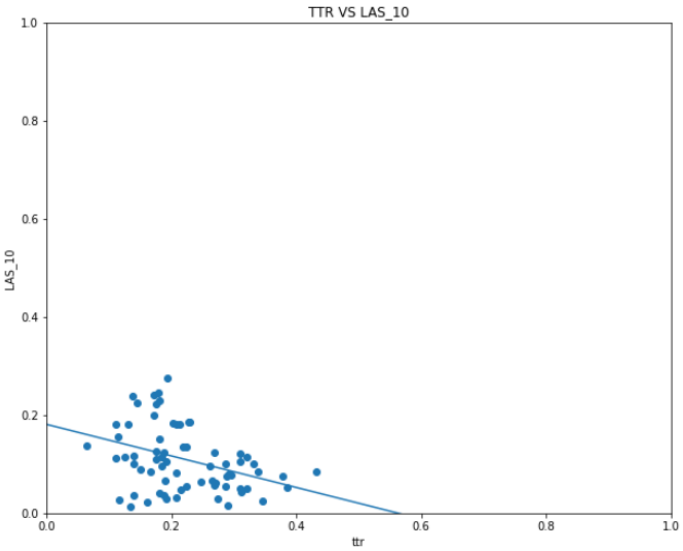
VI. Correlation between metrics

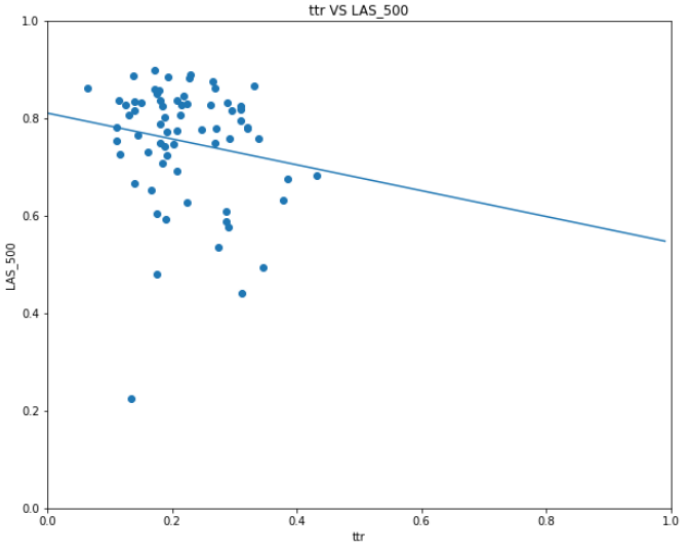
VI.1 Percent of functional words VS LAS for dataset of size 10 (left) and 500 (right)



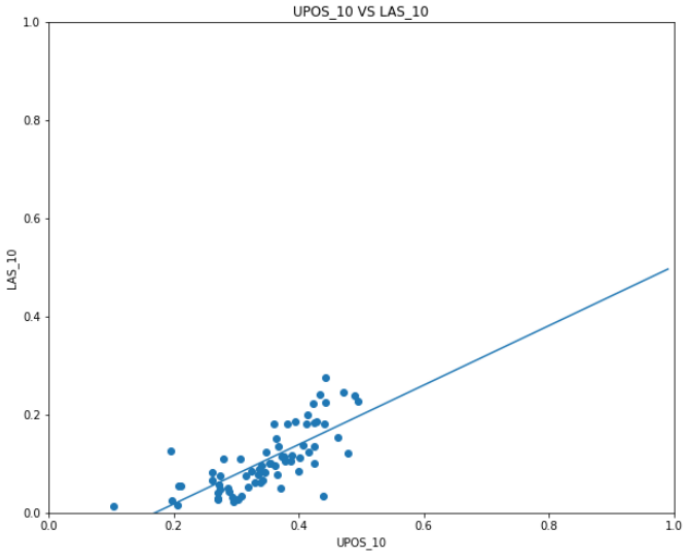


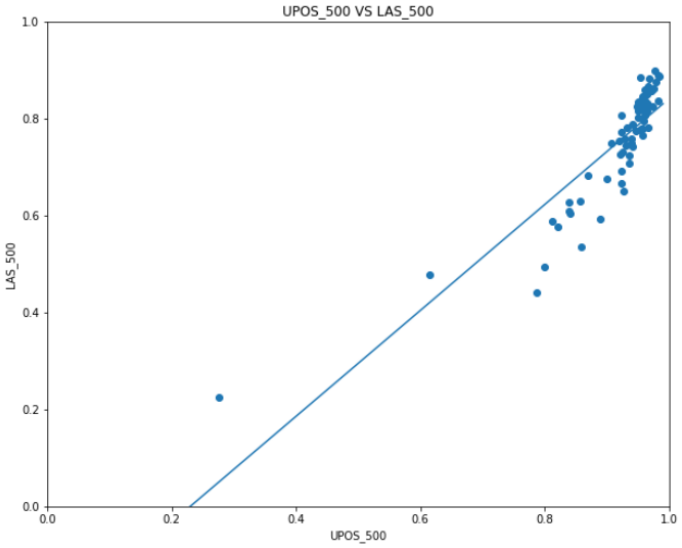
VI.2 Type-token ratio VS LAS for dataset of size 10 (left) and 500 (right)





VI.3 UPOS VS LAS for dataset of size 10 (left) and 500 (right)

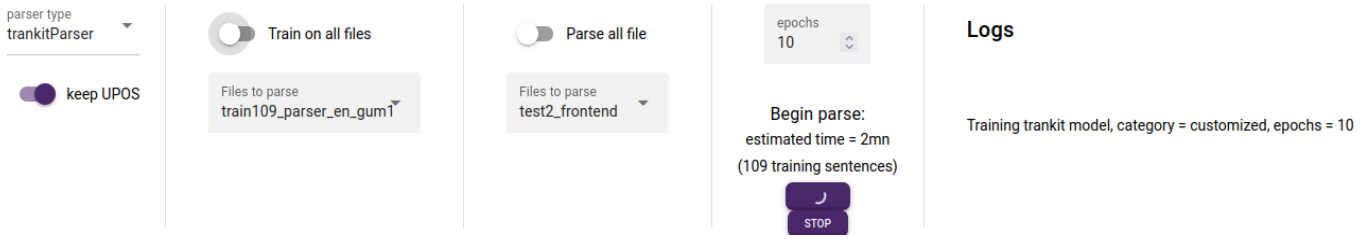




VII. The ArboratorGrew implementation

Screenshot of AboratorGrew's parse options panel:

Users can first choose the gold files as training set, the files to parse, then the parser type such as trankitParser for Trankit and the number of epochs. The keep UPOS option indicates whether the UPOS in selected files to parse need to be kept. If we click the 'begin parse' button, a log message appears to show the current progress, such as data preparation, training and parsing.

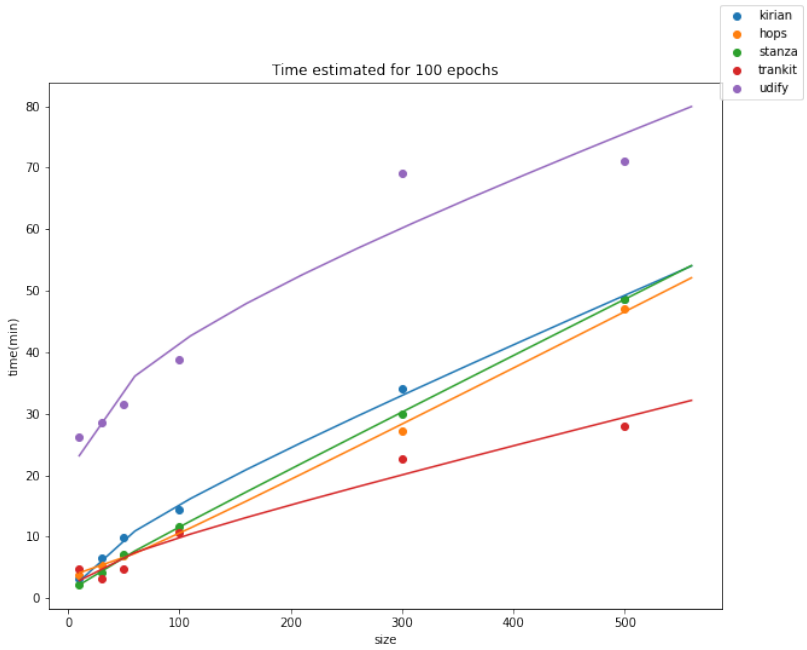


A rough time estimation for each parser:

Empirically, the training and parsing time increases faster than logarithm but slower than a simple line regression, so the logical regression is computed with the following parameters:

$$ftime = A*\log(x + 1) + B*x + C.$$

Note that the effective consumed time may be less than estimated with larger training data.



Réalisations, Obstacles et Perspectives pour l'Outillage du Corse

Laurent Kevers[1] Alice Millour[2]
(1) UMR CNRS 6240 LISA, Università di Corsica - Pasquale Paoli, Avenue Jean Nicoli, 20250 Corte, France
(2) Université Paris 8, 2 rue de la Liberté, 93526 Saint-Denis cedex, France
kevers_l@univ-corse.fr, am@up8.edu

RÉSUMÉ
Présentation des ressources et outils linguistiques développés depuis 2019 pour le corse.

ABSTRACT
Achievements, challenges and perspectives for the tooling up of Corsican.

Presentation of the language resources and tools developed since 2019 for Corsican.

MOTS-CLÉS : langues peu dotées, corse, corpus, ressources linguistiques, outils, TAL.

KEYWORDS: less-resourced languages, Corsican, corpora, lexical resources, tools, NLP.

1 Contexte et objectifs

Le corse, une des 24 langues de France présente sur le territoire métropolitain, est considéré comme « en danger » par l'UNESCO (Moseley, 2010) et est communément repris parmi les langues peu dotées (Leixa *et al.*, 2014; Joshi *et al.*, 2020).

Il s'agit d'une langue présentant une variation dialectale qui s'étend sur quatre, voire cinq aires géographiques (Dalbera-Stefanaggi, 2002, 2007), qui dépassent même les frontières de la Corse en allant jusqu'en Gallura (nord de la Sardaigne). Il existe une intercompréhension entre les locuteurs des diverses aires, celles-ci formant un *continuum* qui se prolonge jusqu'aux variétés centrales et méridionales de l'Italie. Malgré la mise en oeuvre d'une approche polynomique permettant d'englober l'ensemble des variantes dialectales (Marcellesi, 1984), l'écriture de la langue n'est pas standardisée.

Cet article fait le point sur des travaux entrepris depuis 2019 et qui ont pour objectif de constituer un socle de ressources et d'outils accessibles, autant que possible, selon les principes de la science ouverte, et ce afin d'améliorer progressivement le support du corse dans les outils numériques.

Notre ambition est de constituer des ressources et de développer des outils spécialisés pour le TAL [1], mais aussi à destination des apprenants et du grand public.

Cette action, qui est ancrée au sein de la Banque de Données Langue Corse [2] (BDLC), a également pour but l'ajout de fonctionalités au portail BDLC et le développement de modules d'aide au traitement des données linguistiques brutes.

1. https://bdlc.univ-corse.fr/tal/
2. https://bdlc.univ-corse.fr/

Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain *(LIFT)* et Traitement Automatique des Langues *(TAL)*,
Marseille, France, 14 au 15 novembre 2022 , pages 154–161,

2 État de l'art

Depuis 1986, la BDLC recueille, stocke, analyse et restitue des données dialectales relatives aux savoir-faire et aux traditions culturelles corses, par le biais d'enquêtes de terrain en Corse et dans le nord de la Sardaigne. La BDLC contient plus de 2 000 ethnotextes totalisant 317 512 tokens, ainsi que près de 120 000 entrées lexicales contenant la « question » (forme en français), la « réponse » (forme en corse), le « lemme », et la « commune » (localisation).

En dehors du projet BDLC, plusieurs initiatives de différentes natures ont déjà porté sur cette langue peu dotée : le projet Interreg *INTERTESTU* (Chiorboli, 1995) ; le travail mené par l'association ADECEC [3] ; les traducteurs automatiques *Okchakko* [4] ou *Google Translate* [5] ; etc. Un inventaire plus complet des ressources et outils disponibles pour le corse est proposé par Kevers *et al.* (2021).

Les résultats pérennes, diffusés de manière ouverte et directement exploitables pour le TAL étant néanmoins de faible ampleur, l'impulsion décisive permettant la création progressive, cumulative et cohérente des ressources et outils n'a pas eu lieu.

3 Élaboration des ressources et outils pour le TAL corse

3.1 Approche générale

Notre travail vise à améliorer cette situation et s'oriente vers la constitution de corpus corses libres de droits, la création de modules pour le TAL, et enfin la mise à disposition de données et d'outils pour la recherche, ainsi que d'applications à visée pédagogique ou à destination du grand public.

L'approche adoptée s'appuie sur une interaction entre le projet BDLC proprement dit et les développements réalisés dans le cadre du TAL. La BDLC adopte une démarche de terrain qui lui permet de récolter des données linguistiques qui documentent la langue et ses variations, tant en diachronie qu'en synchronie. Les principaux résultats sont constitués par la base de données en tant que telle, par les outils d'accès et de visualisation des données, ainsi que par des publications scientifiques ou de vulgarisation. Les développements en TAL se déroulent à partir de données provenant de diverses sources qui doivent permettre le traitement le plus large et robuste possible de la langue, tout en essayant de respecter un cadre linguistique qui n'est rattaché à aucune norme strictement formalisée. Les résultats concrets sont constitués de ressources linguistiques – entre autres dictionnaires électroniques et corpus – et d'outils.

Notre volonté est que ces deux démarches aboutissent à une interaction et un enrichissent mutuel durant lesquels la BDLC apporte ses données lexicales et textuelles dialectales, ses connaissances linguistiques ainsi qu'un cadre structurant l'automatisation du traitement d'une langue non normalisée. De son côté le TAL propose des outils de traitement, de correction, d'exploration et d'exploitation des données, qu'elles soient brutes ou déjà dépouillées, ce qui ouvre de nouvelles perspectives pour l'accroissement de la base et pour l'enrichissement des connaissances linguistiques.

3. https://www.adecec.net/
4. http://www.okchakko.com/
5. https://translate.google.fr

3.2 Ressources lexicales et variation

Un premier lexique électronique a été extrait de la BDLC. Il comprend 21 108 formes simples ou composées se référant à 12 498 lemmes. Les formes verbales étant sous représentées dans la BDLC, un complément a été créé, en partie automatiquent, en partie manuellement (Kevers & Retali-Medori, 2020; Retali-Medori & Kevers, 2022). La couverture du lexique général a été estimée à 49% des occurrences du corpus d'ethnotextes de la BDLC, et à environ 16% pour les verbes [6]. Ces ressources, qui restent partielles et qui incluent parfois certaines erreurs ou incohérences, n'ont à ce stade pas encore été diffusées.

Du point de vue de la variation dialectale, les ressources lexicales produites par les linguistes de terrain de la BDLC présentent une richesse importante qui demande à être exploitée. Si les premières expérimentations de génération automatique de variantes dialectales sur la base d'entrées dialectales reliées au même lemme – inspirées de Millour & Fort (2019) – sont prometteuses, la méthode pâtit également du manque de cohérence recontré au sein de la BDLC.

L'exploitation du contenu lexical de la BDLC se heurte donc à divers obstacles. En particulier, la catégorie « lemme » pose des difficultés théoriques importantes dans le cadre d'une langue non standardisée. Ce point a fait l'objet de réflexions publiées dans Retali-Medori & Kevers (2022). La multitude d'acteurs ayant participé à renseigner la base au cours des années sans qu'un guide de saisie clair n'ait été imposé en est également une raison. La collaboration entre linguistes de terrain et spécialistes du TAL prend donc tout son sens dans ce projet, la ressource gagnant à être examinée et corrigée semi-automatiquement par des moyens informatiques afin de pouvoir être exploitée ultérieurement à des fins de production de ressources et d'outils pour le TAL.

3.3 Corpus

Dès le départ, nous avons pu disposer du corpus d'ethnotextes de la BDLC (317 512 mots). Celui-ci est représentatif d'un type de textes particulier : des retranscriptions d'entretiens oraux semi-dirigés. Ces textes ne peuvent à eux seuls constituer le substrat nécessaire à l'élaboration des outils de TAL. Dès lors, des corpus de portée plus générale ont été réunis et diffusés (Kevers & Retali-Medori, 2020) : l'encyclopédie collaborative *Wikipedia* en langue corse (919 382 mots), le *blog* journalistique *A Piazzetta* (504 225 mots), ainsi que la traduction corse de la *Bible* (770 560 mots). Ces corpus ont leurs propres caractéristiques. *Wikipedia* contient des documents issus d'un processus d'édition non centralisé qui pose la question de la qualité et la cohérence linguistique de ces textes. Au contraire, *A Piazzetta*, propose un matériau plus uniforme et contrôlé, mais qui intègre également de nombreux commentaires caractéristiques des *blogs*, ceux-ci présentant de fortes variations de la qualité du contenu linguistique, du registre employé, ainsi que des langues utilisées – le français étant fréquemment utilisé à côté du corse. Enfin, la *Bible* reste par nature un corpus très particulier. Par conséquent, la diversification des corpus reste donc un objectif du projet. Une convention conclue avec le réseau Canopé de Corse [7] nous permet de travailler actuellement à un nouveau corpus constitué d'oeuvres littéraires (adulte, jeunesse, enfants) ainsi que de documents relatifs au patrimoine et à l'histoire corses. Ce corpus, qui dépassera les 500 000 mots, sera diffusé très prochainement.

6. Ces chiffres sont sujets à variation en fonction du corpus d'application. Le fait qu'un mot soit reconnu par le dictionnaire ne signifie pas que l'ambiguïté lexicale ait été prise en compte : une forme reconnue par un dictionnaire peut disposer de plusieurs analyses concurrentes, sans que celles-ci soient forcément toujours adéquates pour toutes les occurrences du corpus.

7. https://www.reseau-canope.fr/canope-academie-corse/

Si les corpus sont encore en phase d'élaboration, nous avons d'ores et déjà mis en place une interface de consultation des ethnotextes de la BDLC au travers d'un concordancier [8]. Ce type d'outil permet une exploration potentiellement assez fine des corpus et la visualisation des résultats sous la forme KWIC (*Keywords in Context*). Le corpus peut être filtré selon les méta-données disponibles – thème et localisation – et interrogé au moyen de requêtes simples ou composées, faisant intervenir des formes brutes ainsi que diverses formes d'expressions régulières [9]. L'outil [10] est prévu pour tirer parti des futures annotations qui viendront enrichir nos corpus : identification de langue, parties du discours, lemmes...

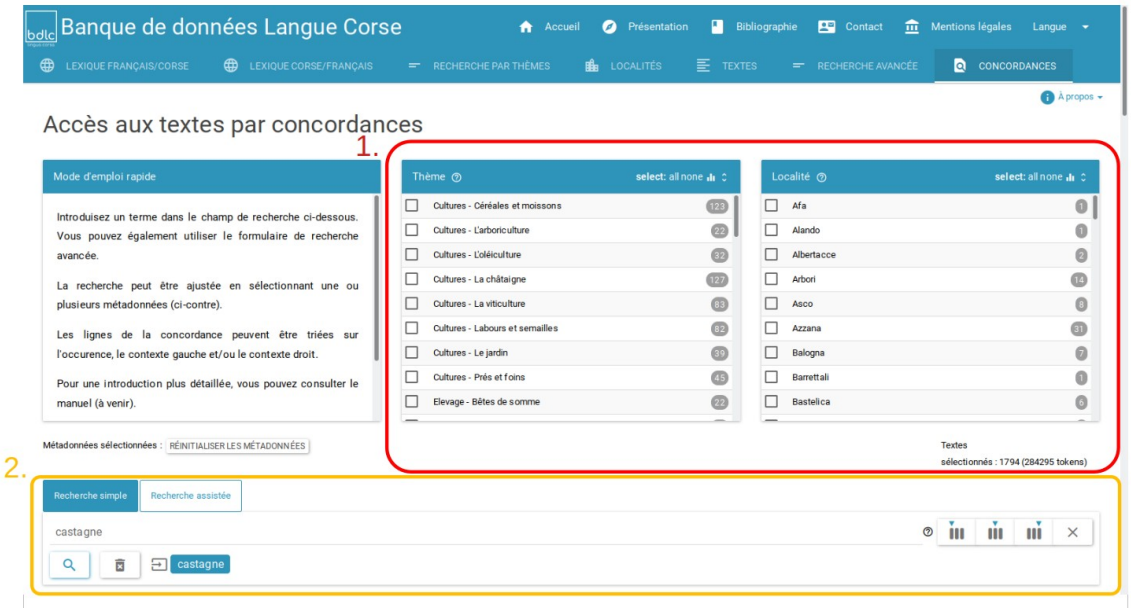


FIGURE 1 – Concordancier : filtrage par méta-données (1.) et zone d'introduction de la requête (2.)

Cette initiative, déjà annoncée dans Kevers & Retali-Medori (2020), constitue un premier enrichissement fonctionnel de la BDLC, à la fois à destination des chercheurs et des enseignants/apprenants. Il est prévu de l'étendre au futur corpus Canopé.

3.4 Outils

Ces différentes ressources ont permis de démarrer l'élaboration de certains modules de TAL, en particulier celui qui concerne l'identification de langue, qui s'avère utile lors de la constitution de corpus. La vérification et l'évaluation de la qualité des textes – comme déjà réalisé pour diverses ressources multilingues de grandes tailles (Kevers, 2022c) – ou l'annotation des documents multilingues [11] (Kevers, 2022a) permettent d'améliorer les corpus. À l'avenir, ce type d'analyse et d'annotation devraient être réalisés sur nos corpus bruts.

D'autre part, grâce à l'annotation manuelle d'un premier corpus de référence de 100 phrases, les premières expérimentations d'annotation morphosyntaxique ont récemment pu être menées. Si nous

8. https://bdlc.univ-corse.fr/concord/
9. Un manuel d'utilisation détaillé a été rédigé et mis à disposition : https://bdlc.univ-corse.fr/concord/docs/user_manual.pdf
10. Adapté et intégré à partir d'un développement réalisé par le Cental (https://uclouvain.be/fr/instituts-recherche/ilc/cental).
11. Spécification de la langue au niveau du mot.

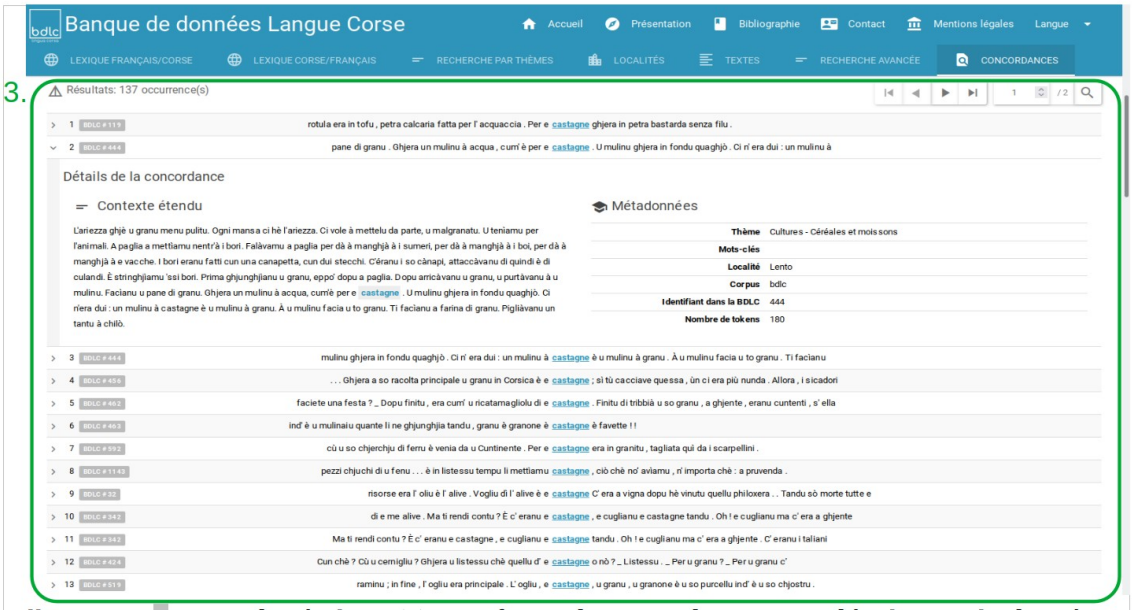


FIGURE 2 – Concordancier : exemple de résultat (3.)

pouvons envisager d'utiliser ce premier modèle d'annotation automatique comme un outil de pré-annotation, l'amélioration des performances reste nécessaire. L'annotation de nouvelles données devrait avoir lieu en 2023 et mener à une progression de la précision de l'outil. Ce module, une fois à maturité, pourra naturellement contribuer à l'annotation semi-automatique de nos corpus, et à leur interrogation au moyen des catégories POS dans le concordancier.

3.5 Questions transversales et méthodologiques

Ce projet est également l'occasion de soulever des questions transversales et communes à différentes langues peu dotées, entre autres la question des droits d'auteurs lors de la constitution de corpus (Kevers & Retali-Medori, 2019), qui a récemment connu un évolution positive avec la transposition française de la directive européenne 2019/790 (Maurel & Rennes, 2021).

L'adoption d'une démarche de science ouverte et le respect des principes FAIR [12] (Wilkinson *et al.*, 2016; Berez-Kroeker *et al.*, 2018) est un point important. Il convient en effet que les résultats des recherches soient reproductibles, que les ressources produites soient disponibles de manière ouverte et pérenne – au minimum pour la recherche – et que les différentes données puissent être correctement identifiées et citées, en particulier au travers des identifiants pérennes (Kevers, 2022b).

Citons aussi les approches par transfert depuis une langue mieux dotée, les méthodes non supervisées ainsi que les démarches participatives de *myriadisation* (Millour, 2020).

Enfin, les aspects méthodologiques liés à la prise en compte des variations dialectales constituent également un point d'intérêt pour de nombreuses langues.

12. Faciles à trouver ; Accessibles ; Interopérables ; Réutilisables

3.6 Obstacles

Au-delà de ces questions, nous pouvons mettre en avant plusieurs difficultés qui ont été rencontrées durant ces dernières années. Tout d'abord, en dépit de la numérisation croissante de la société en général, il reste souvent difficile d'accéder à des contenus de qualité, libres de droits et dans un format structuré natif tel qu'XML. La constitution de corpus reste donc en partie dépendante de problématiques de conversion de formats – par exemple à partir de fichiers PDF – voire de la numérisation de documents imprimés.

D'autre part, nous avons aussi été confrontés à la disponibilité des ressources humaines adéquates. En effet, le projet BDLC s'appuie sur des linguistes spécialistes du corses, mais peu familiers avec les enjeux, méthodes et outils du TAL. La mobilisation de profils spécialisés en TAL n'est de plus pas toujours aisée – en raison de la disponibilité des personnes et des financements – et ceux-ci ne sont pas nécessairement compétents en corse. Enfin, le recrutement de stagiaires est rendu difficile car le cursus universitaire corse ne produit pas d'étudiant disposant de la double compétence linguistique et informatique, et l'intérêt des étudiants de chaque filière pour la discipline complémentaire est limité.

D'une manière générale, la création et la pérennisation de postes dédiés à la problématique des langues peu dotées – et du corse en particulier – sont compliquées, ce qui rend la poursuite des projets à moyen et long termes incertaine.

4 Conclusion

Nous avons résumé les progrès enregistré depuis 2019 et présenté les dernières avancées, en particulier le travail engagé pour la constitution d'un nouveau corpus d'environ 500 000 mots, les premières expérimentations relatives au traitement de la variation ainsi qu'à l'annotation morphosyntaxique.

Les interactions entre linguistique de terrain et TAL ont été mises en évidence, tout comme les différents obstacles auxquels nous avons été confrontés.

Malgré ces difficultés, nous avons progressé sur la production pérenne et ouverte de ressources et outils utiles au traitement automatique du corse. Nous désirons poursuivre ce travail, tout en contribuant le plus possible à une approche générique de l'outillage des langues peu dotées, en particulier durant le projet ANR DiViTal [13] (2022-2025) dont l'Université de Corse est un des partenaires.

Remerciements

Ce travail a été mené grâce au financement CPER : « Un outil linguistique au service de la Corse et des Corses : la Banque de Données Langue Corse (BDLC) » ainsi qu'une bourse post-doctorale de la Collectivité de Corse (CDC).

13. https://divital.gitpages.huma-num.fr/fr/

Références

BEREZ-KROEKER A. L., ANDREASSEN H. N., GAWNE L., HOLTON G., KUNG S. S., PULSIFER P., COLLISTER L. B., THE DATA CITATION AND ATTRIBUTION IN LINGUISTICS GROUP & THE LINGUISTICS DATA INTEREST GROUP (2018). *The Austin Principles of Data Citation in Linguistics*. Rapport interne Version 1.0. https ://site.uit.no/linguisticsdatacitation/austinprinciples.

CHIORBOLI J. (1995). *La gestion du territoire linguistique : INTERTESTU, une base textuelle littéraire et linguistique corse*. Centru di ricerche Corse Gruppulingua, Universuta di Corsica. OCLC : 490793344.

DALBERA-STEFANAGGI M.-J. (2002). *La langue corse*. Volume 3641 de Que sais-je ? Paris : PUF.

DALBERA-STEFANAGGI M.-J. (2007). *Nouvel atlas linguistique et ethnographique de la Corse : Volume 1, Aréologie phonétique, édition revue et corrigée*. Ajaccio : Paris : Comité des travaux historiques et scientifiques - CTHS, Alain Piazzola édition.

JOSHI P., SANTY S., BUDHIRAJA A., BALI K. & CHOUDHURY M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 6282–6293, Online : ACL. DOI : 10.18653/v1/2020.acl-main.560.

KEVERS L. (2022a). CoSwID, a Code Switching Identification Method Suitable for Under-Resourced Languages. In *1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages (SIGUL2022 @LREC2022)*, Proceedings of the the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages (SIGUL2022 @LREC2022), p. 112–121, Marseille, France : European Language Resources Association. Backup Publisher : European Language Resources Association.

KEVERS L. (2022b). L'identifiant pérenne, une clé pour les bases de données linguistiques dans une perspective de science ouverte. In *XXXe Congreso Internacional de Lingüística y Filología Románicas*, La Laguna, Tenerife, Islas Canarias, Spain. HAL : hal-03722878.

KEVERS L. (2022c). L'identification de langue, un outil au service du corse et de l'évaluation des ressources linguistiques. *Traitement Automatique des Langues*, **62**(3). Numéro spécial " Diversité linguistique".

KEVERS L. & RETALI-MEDORI S. (2019). Copyright in the context of tooling up Corsican and other less-resourced languages. In *Proceedings of the 1st International Conference on Language Technologies for All*, p. 198–201, Paris, France : European Language Resources Association (ELRA).

KEVERS L. & RETALI-MEDORI S. (2020). Towards a Corsican Basic Language Resource Kit. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC2020)*, p. 2726–2735, Marseille, France : European Language Resources Association (ELRA).

KEVERS L., RETALI MEDORI S. & TOGNOTTI A. G. (2021). *A Survey of Language Technologies Resources and Tools for Corsican*. Rapport interne, UMR CNRS 6240 LISA, Université de Corse. https://hal.archives-ouvertes.fr/hal-03228733.

LEIXA J., MAPELLI V. & CHOUKRI K. (2014). *Inventaire des ressources linguistiques des langues de France*. Rapport interne, ELDA.

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

MAUREL L. & RENNES S. (2021). La fouille de textes et de données à des fins de recherche : une pratique confirmée et désormais opérationnelle en droit français. https://www.ouvrirlascience.fr/la-fouille-de-textes-et-de-donnees-a-des-fins-de-recherche-une-pratique-confirmee-et-desormais-operationnelle-en-droit-francais.

MILLOUR A. (2020). *Myriadisation de ressources linguistiques pour le traitement automatique de langues non standardisées*. Theses, Sorbonne Universite.

MILLOUR A. & FORT K. (2019). Unsupervised Data Augmentation for Less-Resourced Languages with no Standardized Spelling. In *RANLP*, p. 776 – 784, Varna, Bulgaria. HAL : hal-02280002.

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

RETALI-MEDORI S. & KEVERS L. (2022). La morphologie dans la Banque de Données Langue Corse : bilan et perspectives. *Corpus*, **23**. Numéro thématique « Corpus et données en morphologie ».

WILKINSON M. D., DUMONTIER M., AALBERSBERG I. J. J., APPLETON G., AXTON M., BAAK A., BLOMBERG N., BOITEN J.-W., DA SILVA SANTOS L. B., BOURNE P. E., BOUWMAN J., BROOKES A. J., CLARK T., CROSAS M., DILLO I., DUMON O., EDMUNDS S., EVELO C. T., FINKERS R., GONZALEZ-BELTRAN A., GRAY A. J. G., GROTH P., GOBLE C., GRETHE J. S., HERINGA J., 'T HOEN P. A. C., HOOFT R., KUHN T., KOK R., KOK J., LUSHER S. J., MARTONE M. E., MONS A., PACKER A. L., PERSSON B., ROCCA-SERRA P., ROOS M., VAN SCHAIK R., SANSONE S.-A., SCHULTES E., SENGSTAG T., SLATER T., STRAWN G., SWERTZ M. A., THOMPSON M., VAN DER LEI J., VAN MULLIGEN E., VELTEROP J., WAAGMEESTER A., WITTENBURG P., WOLSTENCROFT K., ZHAO J. & MONS B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, **3**. Article number : 160018 (2016), DOI : 10.1038/sdata.2016.18.

Réseaux de neurones pour une détection automatique des NPI

Ekaterina Kolos[1,3] Pascal Amsili[2,3]
(1) Université Paris Nanterre, 200 Av. de la République, 92000 Nanterre, France
(2) Sorbonne-Nouvelle, 8 avenue de Saint-Mandé, 75012 Paris, France
(3) Laboratoire Lattice, 1 rue Maurice Arnoux, F-92120 Montrouge, France
ekaterina.kolos@sorbonne-nouvelle.fr, pascal.amsili@ens.fr

RÉSUMÉ
Cet article présente un travail en cours sur la détection des NPI pour les textes anglais et présente les résultats obtenus dans la première partie de ce projet. Nous présentons à la fois un corpus annoté, une version préliminaire d'un système d'étiquetage et ses premiers résultats.

ABSTRACT
Automatic Negative Polarity Item Detection

This paper introduces a work in progress on NPI Detection for English texts and presents the results obtained in the first part of this project. We present in this paper an annotated corpus, a preliminary version of a tagging system, and its first results.

MOTS-CLÉS : polarité, monotonie, NPI, FCI, étiquetage de séquences, classification multiclasse.

KEYWORDS: polarity, monotonicity, NPI, FCI, sequence tagging, multiclass classification.

1 Context and motivation

Negative Polarity Items (NPI) are lexical units like the English *any* that are only grammatical in a limited number of contexts, also called licensing contexts. The most common licensing context is negation (1), which explains why they are called NPI, but many other contexts have been shown to licence NPIs, like interrogative sentences, the restrictor of a universal quantifier, or the antecedent clause of a conditional sentence, etc. (Homer, 2020).

(1) (a) *John has *any* friend(s).
(b) John does not have *any* friends.

The list of NPIs (lexical units or constructions) is very heterogeneous, both syntactically (pronouns, determiners, adverbs, NPs...) and semantically. Some NPIs belong to closed classes, which means that we can make a complete list of them (determiners/pronouns...), but other NPIs are built around nouns denoting a small quantity (*a clue*, *a finger*...), so that we have an open list of NPIs. In addition, licensing contexts are also very diverse, and semanticists are still working to establish the list of licensing environments (see Appendix for the list we adopt here), and to try to determine what those environments have in common.

We are concerned in this work with the automatic detection of NPIs and identification of their licensing contexts. Detecting NPIs is not a current task identified as such in the natural language processing

Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain *(LIFT) et* Traitement Automatique des Langues *(TAL)*,
Marseille, France, 14 au 15 novembre 2022 , pages 162–171,

community, however, we consider that it would be interesting to have a way to detect NPIs with a good quality, for several reasons. It would allow linguists to collect data for theoretical investigations on polarity (and also free choice items, see below), as well as monotonicity and negation. It might also prove useful to get additional features for the NLP task of detecting negation and its scope. Downstream applications like natural language generation or text correction may also benefit from a proper identification of polarity items and contexts. Furthermore, NPI detection may be used as a diagnostic classifier (Hupkes *et al.*, 2018) when assessing models' capabilities to learn complex semantic and syntactic concepts (Jumelet & Hupkes, 2018; Jumelet *et al.*, 2021; Bylinina & Tikhonov, 2021).

The task of detecting NPIs is made more complex by the fact that several lexical units that have a NPI function (like *any* or *ever*) can also be used in contexts where they get a different interpretation, dubbed "free choice" (2) (Fauconnier, 1978; Giannakidou, 1998).

(2) You can take *any* book you like.

A separate class of polarity items has been proposed, the so-called free-choice items (FCI), that for some researchers constitute a separate group (Fauconnier, 1978) and for others are a subclass of NPI (Homer, 2020). These items can be licensed, among others, by imperatives, modal verbs, comparatives and superlatives. This class comprises items that behave only as FCI, like *whatever*, but also items which can be used both as FCI and as NPI.

The task we are proposing thus requires not only that NPIs are identified as well as their licensing contexts, but also that they are distinguished from FCI.

Identifying FCI and NPI is not that easy as it might seem at first glance. First of all, these items can be ambiguous with non polar expressions, like (3), where *on earth* is not used as an NPI, *vs.* (4), where the expression is a typical minimizer used as an NPI.

(3) The Arctic is experiencing some of the most rapid and severe climate change *on earth*.

(4) Why *on earth* did you leave me ?

Second, multiple items of different classes can be present in one phrase, e.g. (5), where the first *any* is a weak NPI and the second one an FCI.

(5) Is there *any* specific food I might find in *any* pet shop ?

Finally, minimizers, that can be found alone (4), together with an FCI (6) or an NPI (7), constitute an open class of lexical items.

(6) I would appreciate *any* insight on this *at all*.

(7) They don't know *anything* about wine *at all*.

All these items have a particular set of licensing environments in common, although some of them have more limited distribution than others. We believe it might be of interest to train a system capable to identify those licensing environments in order to :

1. disambiguate known NPI when they are present in a sentence ;
2. identify new NPI that haven't been seen yet ;
3. explore to what extent neural models can grasp complex syntactic and semantic dependencies.

We model the task as a sequence labelling task, with a BIO-scheme. Instead of having only two classes (NPI *vs.* non-NPI) we decided to make a further distinction among NPIs : we separate FCI and NPI, minimizers from the rest of NPIs, and we make a distinction between weak and strong NPIs.

As for the licensing context, we chose to encode the type of context in the tag the sequence receives.

The tags in BIO-scheme follow the pattern `B/I-NPI_Type-Licensor_Type`. This results in 128 possible different B/I-tags, 40 of which were present in the train data.

In the rest of this paper, we present the dataset that we created (§ 2) and we give the first results that were obtained with BiLSTM and BERT (§3). We close the paper with a discussion and perspectives.

2 Annotated dataset

The first stage of our work was to produce an annotated dataset that we could use to train and evaluate our models. We started a pilot annotation campaign, with two annotators, and a rather rich tagset since we not only distinguished FCI from NPI, but also subclasses within NPIs. Sixteen licensing environment classes were distinguished (see Appendix).

We pre-selected a number of NPI candidates – a list that is by no means exhaustive but that could serve as a starting point (see Appendix).

English texts from Universal Dependencies (Nivre *et al.*, 2016) were annotated, making possible further use of syntactic information. A total of 1596 sentences [1] out of a total of 38068 sentences were found to contain one or more of the manually pre-selected NPI candidates and added to the dataset (see Appendix for a detailed table). In case two NPI candidates were present in the sentence, two separate datapoints were created, which resulted in 1734 separate datapoints. [2]

The small percentage of sentences with NPI we found here falls in line with the numbers demonstrated in Jumelet & Hupkes's study (Jumelet & Hupkes, 2018) where a total of 301.836 (2.69%) sentences containing any form of *any* (*anybody, anyone, anymore, anything, anytime*, and *anywhere*) were extracted from 11.213.916 sentences of their Google Books corpus.

We made the following decisions when annotating our data :

1. Minimizers are NPI with the following properties : they cannot be licensed by non-monotonous contexts (c.f. *Exactly two students did anything* vs **Exactly two students lifted a finger to help*) ; they are grammatical in affirmative sentences when negating a negation (c.f. *A : You don't give a damn about my problems. B : But I do give a damn !*) ; they can be licensed by some modal verbs in the sense of irrealis (*You could've lifted a finger to help* vs **You could've done anything to help.*) (Sailer, 2021) ; following Homer, we also add *at all* to this group (Homer, 2020).
2. *Any* is an NPI in negative sentences, when it disappears with the change of polarity. Thus, 8 is an NPI and 9 is an FCI.

1. 2 were removed during the annotation phase
2. In the abstract initially submitted we reported having worked with 1885 data samples ; we eventually reduced the number of items we process during this first stage, eliminating, e.g., *quite* used as NPI in examples like *He's not quite sure about it* != ¬*He's quite sure about it.*

(8) Mary isn't trying *anything* to get Mark back. = ¬ Mary is trying something to get Mark back.

(9) Mary isn't ready to try [just] *anything* to get Mark back. = ¬ Mary is ready to try just anything to get Mark back.

3. *any* is, similarly, an NPI in other negative licensing environments, such as negation in the main clause, implicit negation.
4. *any* is, similarly, an NPI in questions and indirect questions.
5. *any* is, similarly, an NPI in antecedents of conditionals.
6. *any* is, similarly, an NPI when licensed by a restrictor of a universal quantifier.
7. *any* is an NPI when licensed by *only*.
8. *any* and *ever* are FCI when licensed by superlatives and comparatives, as well as *too*-phrases. Having this in mind, we also consider FCI *yet* in *the best I've seen* ***yet***.
9. *any* is an FCI when licensed by imperatives and indirect imperatives.
10. *any* and *ever* are FCI when licensed by restrictors such as relative clauses, which define the set of objects from which one can 'freely choose'.
11. *below*, *before*, *prior to* are considered a separate licensing environment and license NPI and not FCI *any* : *Before he could do* ***anything****, the car crashed into the tree.*
12. the negated *any-... but* is considered an NPI : *I haven't seen* ***anything*** *but care and consideration.*
13. *any* in idiomatic *if any*, *if X is* ***any*** *guide* is annotated as NPI : *John has very few friends, if* ***any*** and is considered to be licensed by the *antecedent of a conditional* environment.
14. *any-* items in *any-... of* are annotated as FCI : *We can't rely on* ***any*** *of them.*
15. *anything*, *whatever* in idiomatic *or* ***anything***, *or* ***whatever*** are annotated as FCI.

We also understand that what we annotate as FCI is not homogeneous. A canonical example of freedom of choice would be, e.g., (10). We also annotated as FCI, however, (11), which literally means '*Every* piece of information will be appreciated' and (12), where *any* is a part of an idiom.

(10) Put in a heater and set it to *anywhere* between 78-82.

(11) *Any* and all information will be appreciated.

(12) She was not having *any* of that.

We consider a more linguistically informed classification, e.g. taking into account semantic properties such as downward-entailment (Ladusaw, 1979), non-/antiveridicality (Giannakidou, 1998; Zwarts, 1998), or Strawson entailment (Von Fintel, 1999) a direction for future work.

3 Experiments

To explore how NPIs can be extracted based on the data we annotated, we preprocess the datapoints, merging, where necessary, multiple tags for one sentence, and learn two different models on the resulting BIO-scheme.

We explore two subtasks : first, the system has to be able to predict whether specific tokens of a sentence form an item of interest for us or not (i.e. identify the NPI's boundaries and disambiguate). Second, we try to predict the item's class : [FCI - weak - strong - minimizer] x 16 licensing environments.

For both purposes, we use two models : one is a **BiLSTM** - a simple architecture of random embeddings followed by a bidirectional LSTM network and a linear layer with a softmax to predict the most probable class for each token. We expect this model to learn licensing environments to the left (*I wonder if anyone has any suggestions*) and to the right (*Anyone have any suggestions ?*) of the NPI candidate. [3] The second model we try is **BERT**, based on a pretrained BERT model from huggingface (`bert-base-uncased`) with a classification unit on top. Our data consists of texts of different genres and domains, and we use the most general BERT model without any domain-specific fine-tuning of the embeddings prior to training the classifier. [4]

For the first subtask, the BiLSTM model correctly predicts NPI boundaries (i.e. the entire sequence of 'B', 'I', 'O' tags) in 88.1% of test sentences, the BERT model - in 96.9% cases. A baseline tagger relying solely on a list of NPIs and matching substrings from it showed 68.1% accuracy on the same data. This metric is later referred to as *acc str* in subtask 2.

Table 1 shows the results for subtask 2 : multiclass classification. To better evaluate the performance of our multiclass classifier we compute the following metrics :

1. *acc str* : estimates the number of entirely correctly predicted sentences ; this is quite strict, since an error in one tag corresponds to an incorrect prediction for the whole sentence, although the sentence might have multiple correctly predicted NPI candidates ;
2. *acc tag* : estimates the number of correctly predicted tokens in the whole test dataset, without taking sentences into account : so, if the test dataset contained 2 sentences each of 3 tokens, these would constitute 6 separate datapoints, each for every token ;
3. *acc bi* : same as *acc tag*, but now only for non-'O' tags, i.e. the 'B' and 'I' tags, where the model had to predict the class of the NPI and its licensor ; we need this metric because the *acc tag* metric is biased due to the large proportion of 'O' tags which are easier to predict ; this third metric only evaluates how many NPI and licensor classes were predicted correctly ;
4. weighted average Precision, Recall and F-Score are also provided.

The quantitative estimation shows clearly that the BERT model outperforms the LSTM pipeline.

model	*acc str*	*acc tag*	*acc bi*	P	R	F1
BiLSTM	0.656	0.982	0.572	0.63	0.62	0.62
`bert-base-uncased`	0.831	0.991	0.787	0.85	0.86	0.83

TABLE 1 – subtask 2 : Multiclass classification results

Apart from the quantitative estimation above, we tried to qualitatively estimate our models by asking them to tag new examples inserted manually. We were particularly interested to know if the models could tag minimizers that they had not previously seen. For example, the minimizer *a hoot*, or minimizers based on swear words never occurred in our training data. In our BERT experiments we

3. A combination of SGD optimizer, cross-entropy loss function, and dropout gave the best results. Bidirectional LSTM proved more capable of learning different licensing contexts.

4. The results we list below were obtained with Adam optimizer, learning rate of 1e-05, 10 epochs.

could identify such previously unseen NPI, as in (13); in other experimental settings BERT only tagged *a* as a *B*-tag (beginning of a minimizer). In any case, the model did not tag *a hoot* in (b) where it is not used as a minimizer, neither did the models consider (c) an example of a minimizer, although it is of similar syntactic structure.

(13) unseen NPI : positive (a) and negative (b, c) examples :

(a)	John	does	not	give	**a**	**hoot**
	O	O	O	O	**B-NPI**	**I-NPI**
(b)	The	owl	gave	a	loud	hoot
	O	O	O	O	O	O
(c)	John	does	not	have	a	cat
	O	O	O	O	O	O

The licensing contexts that were better learnt (first of all, because they were better represented) were negation, direct and indirect questions, comparatives, and antecedents of conditionals (a detailed classification report can be found in the Appendix).

4 Conclusion and Future Work

In this work, we introduced a new annotated dataset for NPI and FCI categorization, as well as a first attempt to categorize these items based on our annotation with the help of deep learning tools.

The models we build seem to be capable of grasping syntactic and semantic information on NPI without any explicit syntactic hints.

The models used here leave room for improvement, for example by adding a CRF layer for consistent *B* and *I* tags, by using semantically aware embeddings or combining BERT and LSTM. Better quality might be achieved by learning the licensor type and the NPI type independently, e.g. with a two-head BERT model.

A further direction of future work would be to annotate a bigger dataset with balanced classes, as well as formalize the annotation guide, invite more annotators and estimate the inter-annotator agreement. The current system could be used to select potential data for this new corpus. One could also explore capabilities of multilingual models, like multilingual BERT, in transferring knowledge of NPI licensing from one language to another, which could prove useful in low-resource scenarios.

Acknowledgments

This work was accomplished during the first author's master internship at the Lattice lab. It was supported by the labex EFL (Empirical Foundations of Linguistics, ANR-10-LABX0083). We would like to thank our intern Thi Hien Linh Nguyen for participating in the first annotation pass, as well as Ilya Kuryanov for assisting in proof-reading and editing the manuscript. We thank the anonymous reviewers for their comments on the first version of this short paper.

Références

BYLININA L. & TIKHONOV A. (2021). Transformers in the loop : Polarity in neural models of language. *arXiv preprint arXiv :2109.03926*.

FAUCONNIER G. (1978). Implication reversal in a natural language. In *Formal semantics and pragmatics for natural languages*, p. 289–301. Springer.

GIANNAKIDOU A. (1998). *Polarity sensitivity as (non) veridical dependency*, volume 23. John Benjamins Publishing.

HOMER V. (2020). *Negative Polarity*, In *The Wiley Blackwell Companion to Semantics*, p. 1–39. John Wiley & Sons, Ltd. DOI : https ://doi.org/10.1002/9781118788516.sem057.

HUPKES D., VELDHOEN S. & ZUIDEMA W. (2018). Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, **61**, 907–926.

JUMELET J., DENIĆ M., SZYMANIK J., HUPKES D. & STEINERT-THRELKELD S. (2021). Language models use monotonicity to assess npi licensing. *arXiv preprint arXiv :2105.13818*.

JUMELET J. & HUPKES D. (2018). Do language models understand anything ? on the ability of lstms to understand negative polarity items. *arXiv preprint arXiv :1808.10627*.

LADUSAW W. A. (1979). *Polarity Sensitivity as Inherent Scope Relations.* The University of Texas at Austin.

NIVRE J., DE MARNEFFE M.-C., GINTER F., GOLDBERG Y., HAJIC J., MANNING C. D., MCDONALD R., PETROV S., PYYSALO S., SILVEIRA N. *et al.* (2016). Universal dependencies v1 : A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 1659–1666.

SAILER M. (2021). Minimizer negative polarity items in non-negative contexts.

VON FINTEL K. (1999). Npi licensing, strawson entailment, and context dependency. *Journal of semantics*, **16**(2), 97–148.

ZWARTS F. (1998). Three types of polarity. In *Plurality and quantification*, p. 177–238. Springer.

Appendix

NPI licensing environments distinguished in this work

1. `1_negation` : negation (negative particles, conjunctions, prepositions (*without*, *against*, *unless*) : *John left home without eating **any** breakfast*, *John won't leave unless he finds **anything** useful*);
2. `2_hidden negation` : hidden negation, i.e. non-affirmative verbs (*I doubt that John ate **any** breakfast*), negative predicates (*unlikely* : *John is unlikely to eat **any** breakfast*), other expressions which we informally identify as having some negative meaning, e.g. *It was a big to-do to find **anyone** who knew it = One couldn't easily find **anyone** who knew it*;
3. `3_quantifier of small quantity` : negative quantifiers, or quantifiers of small quantity (*few/little* : *Few commuters **ever** take the train to work*, *Little can be done to change **anything** for the better*);
4. `4_exactly` : non-monotonous quantifier licensing weak NPI (not found in our data, category reserved for further annotation) : *Exactly two students had **any** success with this task.*
5. `5_question` : questions (*Has **anyone** already figured out the answer ?*);
6. `6_indirect question` : indirect questions (*I wonder if **anyone** already figured out the answer*; *I don't want to comment on whether they did any of that.*) and subjunctives : (*John is sorry that Bill said **anything** against Paul*);
7. `7_antecedent of conditional` : antecedents of conditionals (*If **anyone** notices **anything** unusual, it should be reported*);
8. `8_restrictor of universal quantifier` : restrictors of universal quantifiers (*Every customer who had **ever** purchased anything in the store was contacted*);
9. `9_comparatives and superlatives, too-phrases` : comparatives and superlatives, *too*-phrases, *first*, *last* (*John is taller than **any** other employee*), *John is too short to see **anything***);
10. `10_imperative` : imperatives (*Take **any** book you like*);
11. `11_indirect imperative` : indirect imperatives (*I want you to take **any** book you like*);
12. `12_relative_clause_or_other_restrictor` : relative clauses and other restrictors (*John talked to **any** woman who came up to him*);
13. `13_modal_irrealis` : (*John could have lifted **a finger** to help, but he didn't*); not present in the data, reserved for further annotation;
14. `16_temporal` : a category added later for licensing through *before*, *after*, *prior to* etc (*Before he could do **anything**, the car hit the tree*);
15. `14_other_free_choice` : basically all free-choice usages except those explained by imperatives and explicit restrictors (***Anyone** can do it*);
16. `15_other_NPI` : a category for all other contexts licensing NPI proper, for example, *only* : *Only John brought **any** friends.*

Manually pre-selected NPI Candidates

item	NPI	FCI	ambiguous	was annotated
any	NPI : weak	yes	no	yes
any way	NPI : weak	yes	no	yes
anybody	NPI : weak	yes	no	yes
anyone (any one)	NPI : weak	yes	no	yes
anyhow	NPI : weak	yes	no	yes
anything	NPI : weak	yes	no	yes
anywhere	NPI : weak	yes	no	yes
ever	NPI : weak	yes	yes	yes
either	NPI : strong	no	yes	yes
yet	NPI : strong	yes*	yes	yes
a bean	NPI : minimizer	no	yes	yes
a bit	NPI : minimizer	no	yes	yes
a bite	NPI : minimizer	no	yes	yes
a clue	NPI : minimizer	no	yes	yes
a damn	NPI : minimizer	no	no	yes
a drop	NPI : minimizer	no	yes	yes
a finger	NPI : minimizer	no	yes	yes
a fly	NPI : minimizer	no	yes	yes
a note	NPI : minimizer	no	yes	yes
a penny	NPI : minimizer	no	yes	yes
a single word	NPI : minimizer	no	yes	yes
a thing	NPI : minimizer	no	yes	yes
a word	NPI : minimizer	no	yes	yes
all that	NPI : minimizer	no	yes	yes
an eye	NPI : minimizer	no	yes	yes
an inch	NPI : minimizer	no	yes	yes
at all	NPI : minimizer	no	yes	yes
whatsoever	NPI : minimizer	no	yes	yes
whatever	no	yes	yes	yes
whenever	no	yes	yes	no
wherever	no	yes	yes	no
whoever	no	yes	yes	no
whichever	no	yes	yes	no

TABLE 2 – Our NPI Inventory

Number of sentences with NPI from our list in UD corpora

Corpus	NPI Candidates	Sentences	From Total Sentences
en_ewt-ud-dev.conllu	77	72	2001
en_ewt-ud-train.conllu	699	632	12543
en_ewt-ud-test.conllu	90	86	2077
en_gum-ud-dev.conllu	36	33	843
en_gum-ud-train.conllu	231	216	5660
en_gum-ud-test.conllu	27	26	894
en_atis-ud-dev.conllu	16	16	572
en_atis-ud-train.conllu	99	99	4274
en_atis-ud-test.conllu	12	12	586
en_lines-ud-dev.conllu	65	62	1032
en_lines-ud-train.conllu	169	158	3176
en_lines-ud-test.conllu	62	57	1035
en_partut-ud-dev.conllu	3	3	156
en_partut-ud-train.conllu	103	88	1781
en_partut-ud-test.conllu	13	13	153
en_pronouns-ud-test.conllu	0	0	285
en_pud-ud-test.conllu	25	23	1000

TABLE 3 – Number of sentences with NPI extracted from the Universal Dependencies English Corpora

Classification reports

	LSTM Results				BERT Results			
	P	R	F1	support	P	R	F1	support
NPI_FCI_10	0.00	0.00	0.00	4	0.67	0.50	0.57	4
NPI_FCI_11	0.00	0.00	0.00	2	0.00	0.00	0.00	2
NPI_FCI_12	0.00	0.00	0.00	6	1.00	0.17	0.29	6
NPI_FCI_14	0.54	0.48	0.51	27	0.81	0.96	0.88	27
NPI_FCI_9	0.93	0.78	0.85	18	1.00	1.00	1.00	18
NPI_minimizer_1	0.44	0.50	0.47	8	0.44	0.50	0.47	8
NPI_minimizer_5	0.00	0.00	0.00	1	0.00	0.00	0.00	1
NPI_minimizer_7	0.00	0.00	0.00	1	0.00	0.00	0.00	1
NPI_strong_1	1.00	0.83	0.91	6	1.00	0.83	0.91	6
NPI_strong_5	0.00	0.00	0.00	0	0.00	0.00	0.00	0
NPI_weak_1	0.80	0.77	0.79	31	0.80	0.77	0.79	31
NPI_weak_2	0.33	0.33	0.33	3	0.50	0.33	0.40	3
NPI_weak_3	0.00	0.00	0.00	1	0.00	0.00	0.00	1
NPI_weak_5	0.62	0.84	0.71	19	0.90	1.00	0.95	19
NPI_weak_6	0.33	1.00	0.50	1	1.00	1.00	1.00	1
NPI_weak_7	0.80	0.73	0.76	11	0.91	0.91	0.91	11
micro avg	0.62	0.62	0.62	139	0.88	0.86	0.87	139
macro avg	0.36	0.39	0.36	139	0.63	0.56	0.57	139
weighted avg	0.63	0.62	0.62	139	0.85	0.86	0.83	139

TABLE 4 – Test results of the LSTM model (on the left) the BERT model (on the right)

Structuration automatique en XML d'un dictionnaire électronique de l'indonésien à partir de documents Word

Yaying LIU Damien NOUVEL
Inalco ERTIM
yingya621@gmail.com, damien.nouvel@inalco.fr

RÉSUMÉ
Les dictionnaires électroniques sont de plus en plus utilisés dans le contexte de la diffusion et de la popularisation des appareils électroniques et d'Internet. Dans ce contexte, la numérisation et la structuration des dictionnaires édités pour le format papier a tout intérêt à être réalisée. Le projet que nous présentons a pour objectif de convertir un dictionnaire indonésien, initialement rédigé sous Word, afin d'obtenir des bases de données sous forme de ressource lexicale.

MOTS-CLÉS : Dictionnaire électronique, Ressource lexicale, Indonésien.

KEYWORDS: Electronic dictionary, Lexical resource, Indonesian.

1 Présentation générale

De nos jours, avec les progrès technologiques informatiques et le niveau accru de la demande pour diverses ressources numériques, les usages de ces dernières dans le domaine des humanités numériques sont devenus de plus en plus fréquents et importants. La numérisation et la conversion des dictionnaires a indéniablement un effet positif sur l'enseignement, la diffusion et la préservation des langues. Elles fournissent également une importante source de données pour la recherche dans des domaines tels que la linguistique ou le Traitement Automatique des Langues (TAL). Cependant, un grand nombre de dictionnaires sont actuellement stockés dans des fichiers peu structurés, comme Word, qui permettent aux lexicographes de les éditer mais qui ne sont pas adaptées comme formats de données pour les dictionnaires électroniques (Joffe & De Schryver, 2004). Afin d'obtenir des données adaptées, il faut les convertir en un format structuré.

Il existe plusieurs méthodes de conversion, notamment la saisie manuelle des informations, la conversion à l'aide de langages informatiques et la conversion automatique par apprentissage automatique. Saisir manuellement les informations relatives aux articles qui décrivent les entrées du dictionnaire peut être réalisé avec un logiciel d'édition de dictionnaires, ce qui demande beaucoup de travail mais permet des conversions précises et de bonne qualité. L'approche par le programme informatique à base de règles permet d'économiser en temps de main-d'œuvre, mais ne permet pas d'atteindre un haut niveau de précision de conversion pour tous les articles. Les modèles d'apprentissage automatique, tels que le GROBID-dictionary (Khemakhem *et al.*, 2018), sont encore au stade expérimental, difficiles à mettre en œuvre en pratique.

Ce projet vise à construire un dictionnaire électronique pour le Dictionnaire Indonésien Français général (DIF), à l'aide des langages informatiques et des méthodes de TAL. L'étape actuelle du travail vise à terminer la conversion des ressources du dictionnaire en format Word en une base de données

électroniques valide, vérifiée et utilisable.

2 État de l'art

Dans cette section, nous présentons les différents processus de création d'un dictionnaire électronique. Nous commençons d'abord par la présentation des méthodes de conversion du format Word, puis, nous présentons les différents formalismes de structures des entrées. Pour terminer, nous parlons de la post-édition.

2.1 Conversion au format de stockage

Parmi les projets similaires au nôtre, le DiLAF (Enguehard & Mangeot, 2014) est un projet de dictionnaire dont les sources du dictionnaire ont été saisies sous Word, au format DOC. La première étape consiste à convertir les fichiers en DOCX à l'aide d'OpenOffice, puis à les décompresser pour obtenir un format XML. Ensuite, des des langages réguliers implémentés avec outils (plugin TexFx, Textwrangler, notepad++, etc.) permettent d'obtenir un format LMF (Lexical Mark-up Framework) comme modèle de base d'entrée du dictionnaire. Un autre projet similaire a porté sur un dictionnaire galicien (Guinovart & Simões, 2013). Les premières étapes sont similaires au précédent, afin d'obtenir un format XML. Par la suite, le module Text :RewriteRules de Perl permet de modifier les balises et les contenus du fichier XML. Une fois ces étapes de remplacement et de nettoyage terminées, des DTD sont utilisées pour construire la structure des entrées du dictionnaire. Avec les balises portant des informations textuelles, des expressions régulières sont utilisées pour remplacer les balises inutiles par de nouvelles balises.

Contrairement aux méthodes choisies pour ces deux projets, nous avons choisi une méthode de conversion différente pour notre projet. Tout d'abord, nous avons utilisé Python pour convertir les fichiers DOC en fichiers DOCX, puis nous avons utilisé le module Python « DOCX-Python », pour repérer les différentes polices et caractéristiques structurelles du contenus des entrées. Nous avons ensuite extrait directement les différentes structures des entrées qui sont stockées en mémoire selon une structure adaptée aux contenus extraits. Enfin, nous avons utilisé le module « Yattag » pour exporter ces contenus avec des balises correspondant aux différentes parties des entrées, afin de générer un fichier XML pour chaque fichier Word.

2.2 Formalismes de structuration d'entrées lexicographiques

Nous adoptons la norme TEI (Sperberg-McQueen *et al.*, 1994) qui est une normalisation bien établie qui s'est avérée populaire au sein de la communauté lexicographique. Les principaux éléments structurels du chapitre des dictionnaires TEI sont présentés dans la description de (Romary, 2013) :

— `<entry>` est l'élément structurant de base d'un lexique et regroupe les informations de forme, les informations grammaticales, les informations de sens et les renvois ;
— `<form>` peut être utilisé pour décrire une ou plusieurs formes associées à une entrée ;
— `<gramGrp>` regroupe toutes les caractéristiques grammaticales qui peuvent être attachées à l'entrée dans son ensemble (par le biais de son appartenance à la classe de modèle mo-

del.entryPart.top), à une forme spécifique (à travers la classe de modèle model.formPart) ou encore comme contrainte sur l'un des sens d'un mot (toujours à travers model.entryPart.top) ;
— `<sense>` rassemble toutes les informations relatives aux sens, c'est-à-dire les définitions, les exemples, l'usage et les informations d'utilisation et les notes supplémentaires.

TEI n'établit pas de modèle de base pour les entrées. Les quatre structures de base proposées par (Romary, 2013) sont basées sur l'analyse et le résumé de la structure de base de l'entrée. De plus, TEI permet de choisir entre différentes options d'encodage pour le même élément d'information, le lexique encodé dans TEI peut alors avoir des schémas différents.

D'autres formats lexicographiques existent, comme les structures LMF (Francopoulo *et al.*, 2006) et OntoLex-Lemon (McCrae *et al.*, 2017), qui proposent un modèle général comprenant le modèle de base générique et des extensions de ce modèle de base. Toutefois, les extensions de LMF doivent s'appuyer sur la partie *core* pour décrire les données, ce qui n'est pas explicitement souligné par le modèle OntoLex-Lemon. Diverses extensions de LMF mettent l'accent sur l'adaptation aux données du domaine TAL, tandis que OntoLex-Lemon met l'accent sur certaines caractéristiques des entrées dans le domaine des ontologies. Par exemple, dans le modèle OntoLex-Lemon le sens actuel d'un mot est donné par référence à un concept ontologique et les sens lexicaux ne représentent que la mise en correspondance d'un mot à un concept. LMF est moins flexible que TEI. Il permet plus de contrôle sur les pratiques d'encodage et fournit un formalisme plus contraignant pour la modélisation de l'information lexicale.

2.3 Post-édition

À l'heure actuelle, la façon dont est envisagée la création des bases de données pour les dictionnaires électroniques est d'utiliser un langage informatique pour créer et alimenter la base de données, mais cette méthode peut générer des contenus erronés, que l'ordinateur ne pourra pas traiter correctement. Il existe plusieurs manières de réaliser la post-édition. Par exemple, Jürviste *et al.* (2011) ont choisi EELex comme éditeur pour la post-édition de projet. Leur dictionnaire, EBD (« Basic Estonian Dictionary »), est un système interactif, conçu pour les apprenants estoniens (Kallas *et al.*, 2015). Enguehard & Mangeot (2014) ont choisi la plateforme Jibiki comme éditeur pour la poste-édition. Jibiki (Mangeot, 2002) est basé sur un modèle d'interface HTML instancié par l'entrée lexicale à publier. Au lieu de choisir directement des logiciels d'édition existants, Simões *et al.* (2016) ont choisi de créer leurs propres programmes de post-édition. Ils ont choisi eXiste-BD pour importer leur XML.

Pour notre projet, nous utilisons une autre plateforme d'édition du dictionnaire, Lexonomy (Měchura *et al.*, 2017). Cette plateforme open-source permet la rédaction et la publication de dictionnaires.

3 Description de l'indonésien et des sources

Comme la majorité des langues austronésiennes, l'indonésien est une langue agglutinante. L'une des caractéristiques les plus importantes est que la fonction grammaticale est exprimée par l'ajout de différents affixes au début ou à la fin des noms, des verbes, etc., ces informations étant présentes dans le dictionnaire. Dans cette langue, il y a trois grandes catégories morphologiques : les mots de base, les mot-outils et les affixes. Les mots de base peuvent fonctionner de façon autonome, mais

sont également susceptibles d'affixation. Ils peuvent aussi être redoublés (Sneddon *et al.*, 1996).

Le Dictionnaire indonésien français général (Labrousse, 1984) a été réalisé par Pierre Labrousse, professeur d'indonésien à l'Inalco. Le projet de numérisation du DIF a été initié par l'Inalco pour faire évoluer le dictionnaire papier en un dictionnaire électronique en ligne, afin de faciliter la recherche des entrées et le partage de cette ressource. Après cette édition papier du dictionnaire, P. Labrousse a décidé d'en réaliser une version électronique. Pour cela, il a rédigé les articles de cette nouvelle version du dictionnaire au format DOC, et l'a organisé en deux niveaux de structure : le niveau sémantique et le niveau lexical. Au niveau sémantique, les fichiers sont ordonnés par deux grandes sujets : « l'homme » et « la société ». Chaque grand sujet est organisé en des sous-sujets. Par exemple, dans le sujet « l'homme », il existe plusieurs sous-sujets tels que *usia* (l'âge), *tubuh* (le corps), *perasaan* (la perception), etc. Dans chaque fichier, les articles du même sujet sémantique sont ordonnés alphabétiquement.

Un article complet se compose de deux lignes, dont la première contient l'entrée, et souvent des informations historiques. La deuxième ligne contient toutes les informations relatives à l'entrée, notamment les formes des mots, les définitions, les exemples, les traductions, etc. Les différentes parties d'un article sont codées par utilisation de symboles, de polices, de formats.

@II. AZAM. [ar.]
n. intention f., propos m. vv. *niat, maksud, tujuan.* *** berazam. 1.** avoir une intention : *berazam berdikari* avoir l'intention d'être indépendant. vv. *berniat, bermaksud.* **2.** être résolu, déterminé. vv. *bertekad.* *** mengazamkan.** avoir l'intention de, faire tout ce qui est en son pouvoir pour : *mengazamkan hidupnya untuk berdakwah* consacrer sa vie à la propagation de la foi. $ *mengazamkan diri* se consacrer. vv. *meniatkan, memaksudkan.* *** keazaman.** volonté f., intention f. vv. *niat, maksud.*

FIGURE 1 – Exemple d'article de dictionnaire DIF

La figure 2 montre un article général du dictionnaire. Dans la première ligne, il y a deux blocs. Un bloc commence par le symbole « @ », suivi de la forme associée à l'entrée (mot vedette), toujours en gras et en lettres capitales. Le deuxième bloc est entre crochets. Dans la deuxième ligne, nous trouvons tour à tour des informations sur la catégorie, la définition, l'exemple, la traduction, la référence, les synonymes et des sous-entrées. Ces informations apparaissent en même temps que certains symboles ou chiffres spéciaux (« * », « vv. », « $ », etc.). De nombreux échanges avec l'auteur du dictionnaire ont permis de déterminer les parties d'un article.

4 Déroulement du projet

Dans cette section, nous présentons deux grandes parties. Dans la première partie, « Conversions », nous allons décrire les méthodes de deux différentes conversions. Dans la deuxième partie, nous présenterons le travail de post-édition avec le logiciel d'éditeur Lexonomy.

4.1 Conversions

Conversion DOC en DOCX

Constatant que les structures du fichier original sont marquées par leur propre format, nous utilisons le module Python-DOCX de Python qui peut traiter le contenu du fichier DOCX en fonction des différentes caractéristiques de format. Il suffit de convertir au préalable le fichier DOC en fichier DOCX, puis d'analyser ce fichier DOCX. Avec le module *pywin32* de Python, nous avons converti les fichiers DOC en DOCX.

Conversion DOCX en XML

Comme il s'agit d'une étape cruciale de tout le projet, nous décrivons en détail le processus de conversion des articles. La première étape de ce processus est l'extraction et la deuxième étape est l'ajout de balises XML. La logique de ce processus est d'extraire séparément les différents éléments des entrées et d'y ajouter les balises appropriées, les formats ainsi en des entrées qui répondent aux demandes du stockage électronique.

Extraction

Après avoir obtenu le fichier DOCX nécessaire pour la session d'extraction, nous analysons la structure des articles afin d'implémenter un algorithme, sous forme d'un programme Python. Nous avons choisi une méthode d'extraction qui peut extraire et créer une structure assez générique et complète. Elle comprend les contenus tels que l'entrée, la fenêtre historique, plusieurs définitions, les exemples indonésiens, leurs traductions françaises, la partie de forme affixée et la partie de mot composé, comme le montre la figure 2.

Dans cet exemple, nous divisons l'article en six parties : l'entrée, la fenêtre historique, la catégorie, les définitions, la partie de forme affixée et la partie de mot composé.

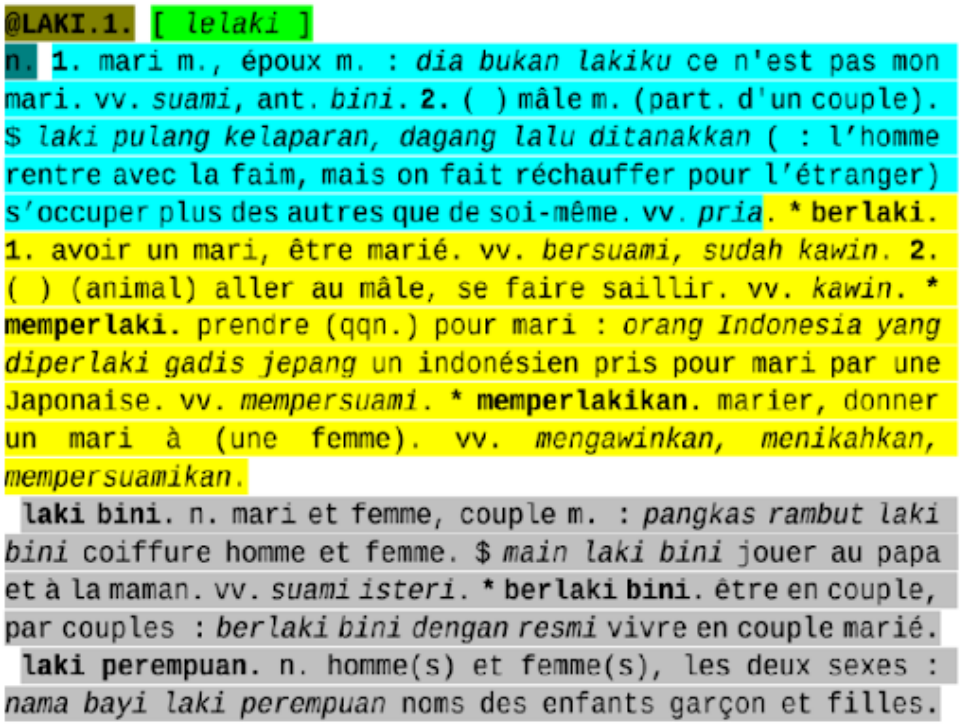


FIGURE 2 – Exemple d'article divisé en six parties

Algorithme du programme d'extraction

La logique de l'algorithme du programme d'extraction est basée sur les informations relatives à la structure d'article, construite en utilisant différents symboles spéciaux et formats textuels, illustrée

par le schéma 3. Au préalable, le programme doit éviter les renvois, qui commencent par « @ » et « < > », dans ce cas le programme n'en tient pas compte. Sinon, si la partie entrée contient le symbole « @ » (sans chevrons) et des lettres en gras (run.bold pour Python-DOCX), le programme peut démarrer l'extraction d'une entrée lexicale.

La figure 2 présente la structure principale d'une entrée, qui contiendra au maximum six éléments à extraire. Le premier élément est le mot vedette, en majuscules. Nous déterminons ensuite si l'entrée contient une fenêtre historique, en testant la présence de crochets « [] », et en utilisant le cas échéant une expression régulière pour la capturer. Ensuite, le programme recherche la présence d'une catégorie lexicale à l'aide d'une liste des catégories lexicales possibles. Une entrée commence par une définition, un exemple en italique et sa traduction française au format *normal*, ces parties sont enregistrées. De même, il repère ensuite les affixes, en gras et suivis d'une étoile « * », les mots composés, en gras mais sans « * ». Notons qu'une entrée peut contenir plusieurs définitions (et traductions associées), qui sont alors numérotées. Toutes ces caractéristiques de format nous permettent de les détecter dans une très large majorité des cas. À la fin de chaque définition peuvent apparaître la référence, le synonyme et l'expression composée, qui sont indiqués respectivement par les signes « vv. », « # », et « ♢ » repérées et extraites par langages réguliers.

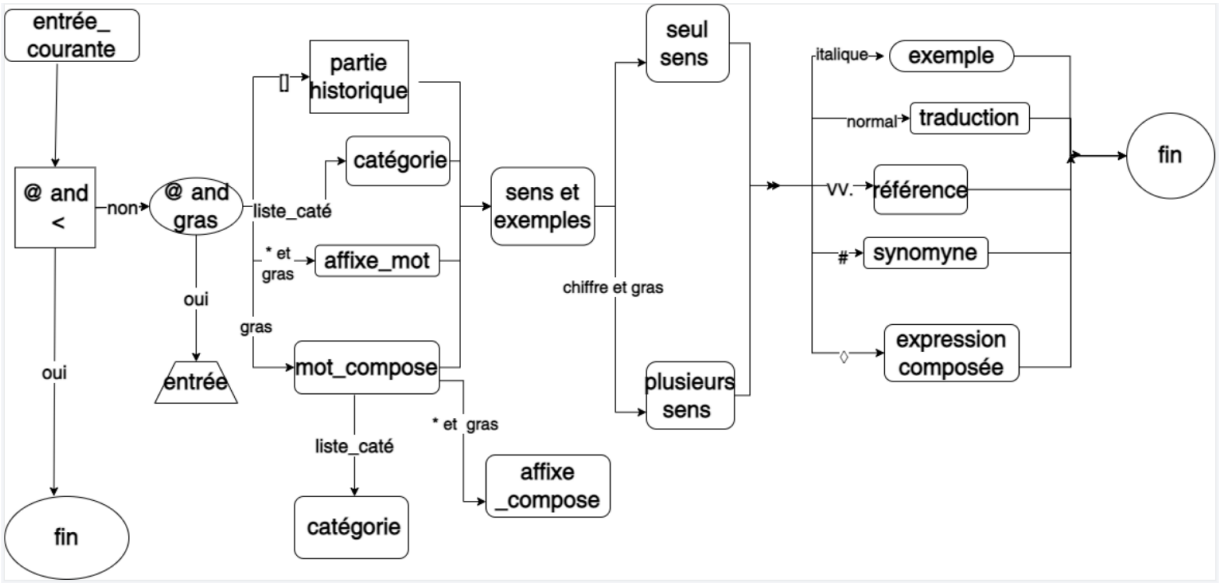


FIGURE 3 – Schéma d'extraction

Génération du format XML

Dans notre processus d'extraction de la structure, tous les articles sont extraits dans des dictionnaires en mémoire afin de produire des entrées au format XML. Nous utilisons la librairie Python Yattag pour faciliter la génération du XML et rendre plus lisible l'organisation de la structure du fichier en sortie. La figure 4 montre un exemple d'article dans un fichier XML.

```
<sense n="1">
    <def> placenta m. :</def>
    <exemples>
        <cite type="exemple" xml:lang="id">
            <quote>penguburan ari-ari</quote>
        </cite>
        <cite type="translation" xml:lang="fr">
            <quote>enterrement du placenta. $</quote>
        </cite>
        <cite type="exemple" xml:lang="id">
            <quote>(exp)(tali) ariarikemudian ariari dipotong</quote>
        </cite>
        <cite type="translation" xml:lang="fr">
            <quote>cordon ombilical m. :ensuite on coupe le cordon ombilical. vv..</quote>
        </cite>
    </exemples>
    <xr type="reference">
        <ref>tembuni, bali, uri, plasénta</ref>
    </xr>
    <xr type="syn">
        <ref></ref>
    </xr>
    <xr type="exp_compose">
        <ref></ref>
    </xr>
</sense>
```

FIGURE 4 – Une partie d'entrée sous TEI dans un fichier XML

4.2 Lexonomy (post-édition)

Lexonomy [1] est une plateforme en ligne pour l'écriture et la publication de dictionnaires. Sa mission est de fournir un outil facile à utiliser pour les petits et moyens projets de dictionnaires. Sur cette plateforme, nous pouvons mettre en œuvre les trois principales fonctions d'importation, de modification et d'exportation. La figure 5 montre l'interface de la plateforme pour modifier une entrée, qui permet déditer, d'ajouter ou de supprimer des nœuds XML avec une interface agréable à utiliser et facile à prendre en main.

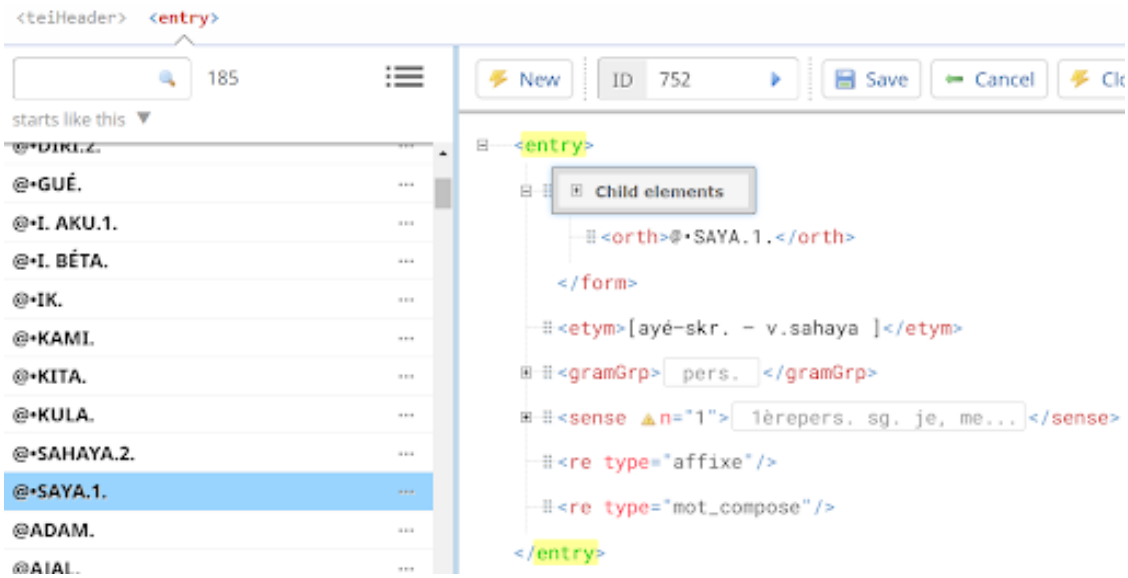


FIGURE 5 – Capture d'interface de la modification de Lexonomy

1. https ://www.lexonomy.eu/

5 Résultats et discussion

Bien entendu, le résultat final du projet dépend directement de la précision de la partie extraction, puisque la partie ultérieure du processus d'ajout de balises est entièrement basée sur les balises du dictionnaire extraites dans les contenus des entrées formatées sous Word. Le modèle d'extraction existant est un modèle générique et implémenté incrémentalement, qui a été construit pour la plupart des structures d'entrées.

Sur un échantillon de 329 articles du DIF, nous calculons que notre programme peut traiter 89,04% des articles, ce qui nous épargne beaucoup de travail manuel. Analysons maintenant le fichier XML final obtenu par ce processus de conversion.

Pour vérifier l'extraction, 16 fichiers sur deux sujets distincts ont été vérifiés manuellement. Le sujet « homme » produit 6 fichiers XML contenant 118 entrées lexicales. Le sujet « âge » génère 10 fichiers XML, qui contiennent 157 entrées extraites correctement par le programme (dont certaines résultent de mots composés et des formes affixées, traitées comme des entrées individuelles).

L'analyse de ces résultats nous a permis de conclure que le programme existant était capable d'extraire une grande partie du contenu automatiquement, avec une précision satisfaisante. Concernant les erreurs, deux raisons principales ont été identifiées. La première concerne les erreurs humaines liées à une utilisation erronée des formats dans le fichier source (par exemple des mots composés précédés de « @ », ce qui n'est pas prévu. La seconde est liée à l'incomplétude du programme : comme il n'est pas possible d'analyser la structure de toutes les entrées, en particulier la partie concernant sens, le programme existant est incomplet et n'est pas en mesure d'extraire le contenu avec précision face à certaines structures trop spécifiques et inattendues.

Bien que le fichier XML existant comporte quelques erreurs et nécessite une session de relecture ultérieure, le mode de conversion automatique réduit beaucoup le travail humain. En plus, dans ce mode de transformation détourné, toutes les entrées sont décomposées et reconstruites, ce qui facilite leur utilisation sous forme électronique (consultation, site web, outils TAL, etc.). Parce que tous les contenus requis de l'article sont extraits et stockés séparément, les lexicographes pourront alors éditer les entrées et reconstruire la structure des articles, sans tenir compte des contraintes de mise en page et de lisibilité des dictionnaires papier.

6 Conclusion

Le travail décrit dans cet article porte sur la conversion d'un dictionnaire de l'indonésien depuis sont format actuel sous Word vers un format XML structuré et exploitable. Pour ce faire, un programme de conversion a été implémenté, qui s'appuie sur les formats (symboles, gras, italique, etc.) du fichier source afin d'extraire les informations utiles en mémoire et de produire un fichier XML contenant les entrées structurées selon un format TEI. Le programme donne des résultats satisfaisants pour une tâche qui serait autrement fastidieuse. Les fichiers XML résultants sont déposés sur une plateforme d'édition de dictionnaire, Lexonomy, qui facilitera leur post-édition par les auteurs de dictionnaires et pourra permettre de publier le dictionnaire et de rechercher dans ses entrées.

Références

ENGUEHARD C. & MANGEOT M. (2014). Computerization of african languages-french dictionaries. *arXiv preprint arXiv :1405.5893*.

FRANCOPOULO G., GEORGE M., CALZOLARI N., MONACHINI M., BEL N., PET M. & SORIA C. (2006). Lexical markup framework (lmf). In *International Conference on Language Resources and Evaluation-LREC 2006*.

GUINOVART X. G. & SIMÕES A. (2013). Retreading dictionaries for the 21st century. In *2nd Symposium on languages, applications and technologies* : Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

JOFFE D. & DE SCHRYVER G.-M. (2004). Tshwanelex : A state-of-the-art dictionary compilation program. In *11th EURALEX International Congress (EURALEX-2004)*, p. 99–104 : Faculté des Lettres et des Sciences Humaines.

JÜRVISTE M., KALLAS J., LANGEMETS M., TUULIK M. & VIKS Ü. (2011). Extending the functions of the eelex dictionary writing system using the example of the basic estonian dictionary. *eLexicography in the 21st Century : New Applications for New Users, Proceedings of eLex*, p. 106–112.

KALLAS J., KILGARRIFF A., KOPPEL K., KUDRITSKI E., LANGEMETS M., MICHELFEIT J., TUULIK M. & VIKS Ü. (2015). Automatic generation of the estonian collocations dictionary database. In *Electronic lexicography in the 21st century : linking lexical data in the digital age. Proceedings of the eLex 2015 conference*, p. 11–13.

KHEMAKHEM M., HEROLD A. & ROMARY L. (2018). Enhancing usability for automatically structuring digitised dictionaries. In *GLOBALEX workshop at LREC 2018*.

LABROUSSE P. (1984). *Dictionnaire général : Indonésien-français*. Cahiers d'Archipel. Éditions de la Maison des sciences de l'homme, Paris.

MANGEOT M. (2002). Projet papillon : intégration de dictionnaires existants et gestion des contributions. In *JST'02 Journées Science et Technologie*, p. 64–65.

MCCRAE J. P., BOSQUE-GIL J., GRACIA J., BUITELAAR P. & CIMIANO P. (2017). The ontolex-lemon model : development and applications. In *Proceedings of eLex 2017 conference*, p. 19–21.

MĚCHURA M. B. *et al.* (2017). Introducing lexonomy : an open-source dictionary writing and publishing system. In *Electronic Lexicography in the 21st Century : Lexicography from Scratch. Proceedings of the eLex 2017 conference*, p. 19–21.

ROMARY L. (2013). Tei and lmf crosswalks. *arXiv preprint arXiv :1301.2444*.

SIMÕES A., ALMEIDA J. J. & SALGADO A. (2016). Building a dictionary using xml technology. In *5th Symposium on Languages, Applications and Technologies (SLATE'16)* : Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

SNEDDON J., AUSTRALIA. COMMONWEALTH DEPT. OF EMPLOYMENT E., TRAINING & EWING M. (1996). *Indonesian : A Comprehensive Grammar*. Comprehensive grammars. Routledge.

SPERBERG-MCQUEEN C., FOR COMPUTERS A. & THE HUMANITIES (1994). *Guidelines for Electronic Text Encoding and Interchange*. Electronic book library. Text Incoding Initiative.

Etude de la Compréhension Spatiale Multimodale des Modèles Transformers Vision-Langage

Emmanuelle Salin[1]
(1) Laboratoire d'informatique et systèmes, 163 avenue de Luminy, 13288, Marseille cedex 09, France
emmanuelle.salin@lis-lab.fr

RÉSUMÉ

Plusieurs études ont montré que les modèles Transformers Vision-Langage ont du mal à appréhender des concepts spatiaux à un niveau multimodal. En utilisant un jeu de données inspiré de CLEVR, nous avons sondé les performances des modèles UNITER, LXMERT et ViLT sur plusieurs tâches de compréhension spatiale multimodale, afin de comprendre les facteurs du pré-entraînement impactant leurs performances.

ABSTRACT

Study of the Multimodal Spatial Understanding of Vision-Language Transformers Models

Several studies have shown that Vision-Language Transformer models have difficulty understanding spatial concepts at a multimodal level. Using a CLEVR-inspired dataset, we probed the performance of UNITER, LXMERT and ViLT on several multimodal spatial comprehension tasks, in order to understand what part of the pre-training impacts their performance.

MOTS-CLÉS : Langage, Vision, Multimodalité.

KEYWORDS: Language, Vision, Multimodality.

1 Introduction

Les modèles Vision-Langage basés sur l'architecture Transformers apprennent à partir de larges jeux de données constitués d'image et de leurs descriptions textuelles à créer des représentations multimodales jointes texte images. Ils atteignent de très bonnes performances dans de nombreuses tâches multimodales, comme des tâches de raisonnement multimodales (Suhr *et al.*, 2018) ou de questions sur des images (Goyal *et al.*, 2017).

Cependant, plusieurs études (Salin *et al.*, 2022), (Rösch & Libovickỳ, 2022) montrent que le pré-entraînement de ces modèles ne leur permet pas de réussir les tâches de raisonnement spatial multimodal. En effet, dans (Salin *et al.*, 2022), nous avons sondé les modèles pour mieux analyser leurs capacités textuelles, visuelles et multimodales sur des jeux de données constitués à partir d'images réelles. Nous avons montré que les modèles parviennent à comprendre le concept de couleur à un niveau multimodal, alors qu'ils ne parviennent pas à associer les informations visuelles aux mots qualifiant la position et de taille d'objets dans l'image. L'utilisation d'un jeu de données réelles limite cependant les possibilités d'évaluation des performances des modèles, car les annotations manquent de précision et peuvent être biaisées.

Nous voulons étudier plus précisément les différences de compréhension multimodale de la position

en fonction des modèles. Pour cela, nous construisons un nouveau jeux de données avec des images synthétiques inspiré de CLEVR (Johnson *et al.*, 2017) pour permettre une évaluation des modèles Transformers Vision-Langage, à partir duquel nous établissons plusieurs tâches multimodales. Nous évaluons les modèles UNITER (Chen *et al.*, 2020), ViLT (Kim *et al.*, 2021) et LXMERT (Tan & Bansal, 2019) sur ces tâches afin de vérifier leur compréhension multimodale de la position d'objet. Nous cherchons notamment à savoir à quel point ces modèles ancrent leur compréhension du texte dans l'image.

1.1 Compréhension de la taille :

Les données image/texte utilisées pour le pré-entraînement des modèles vision-langage ne leur permettent pas d'obtenir une bonne compréhension multimodale de la taille. Une étude des différents jeux de données utilisés nous montre que les adjectifs qualifiant la taille des objets sont mal équilibrés. En effet, les annotateurs écrivent souvent des références à la taille d'un objet :

— Si la taille de l'objet dans l'image est particulièrement inhabituelle pour l'objet considéré (ex : "un petit avion"). Les annotateurs ne décriront pas souvent la taille des mêmes objets quand celle-ci est considérée comme "normale".

— Si elle fait partie d'une expression couramment utilisée, ou pour insister sur la taille de l'objet (ex : "un grand gratte-ciel", "un petit chaton"). L'adjectif qualificatif opposé est alors peu utilisé avec ces objets. L'annotateur préférera dans ce cas décrire les objets avec un autre mot (ex : "un chat" plutôt que "un grand chaton").

La compréhension de la taille par le modèle se heurte donc à la subjectivité de l'annotation, et au mauvais équilibre des mots décrivant la taille dans les données, qui conduisent les modèles Vision-Langage à trop reposer sur le biais textuel.

1.2 Compréhension de la position :

Le concept de position est moins impacté par le mauvais équilibre des données, notamment pour l'utilisation des mots "gauche" et "droite". Les annotateurs ne vont pas préférer utiliser l'un des deux en fonction de l'objet considéré ou du contexte. Nous voulons donc savoir ce qui peut expliquer le mauvais ancrage multimodal de la position. Différents aspects des modèles Vision-Langage peuvent affecter leur compréhension, comme les représentations visuelles utilisées en entrée des modèles, l'architecture Transformers et notamment les encodages de position, les tâches de pré-entraînement ou les jeux de données utilisés.

2 Méthodologie

Nous étudions la compréhension spatiale multimodale des modèles LXMERT, ViLT et UNITER, en utilisant un jeu de données formé d'images synthétiques inspiré de CLEVR. Chaque image est formée de un ou deux objets simples comme visible sur la Figure 1. Les objets sont composés de plusieurs attributs (couleur, taille, forme) qui peuvent être utilisés pour les décrire dans le texte et les distinguer des autres objets dans l'image. Nous évaluons la compréhension multimodale de ces attributs sur des images synthétiques pour les modèles UNITER, LXMERT et ViLT. Le concept de couleur est celui que les modèles arrivent le mieux à appréhender. Nous montrons dans le tableau 1 les résultats des

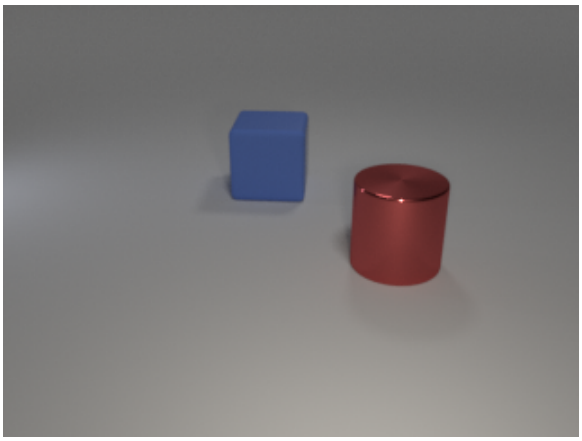


FIGURE 1 – Exemple d'image synthétique contenant deux objets

Modèle	UNITER	LXMERT	ViLT
Accuracy	88.4	80.0	91.1

TABLE 1 – Évaluation de UNITER, LXMERT et ViLT sur une tâche de compréhension multimodale de la couleur réalisée sur un dataset d'images synthétiques inspiré de CLEVR

modèles sur une tâche de compréhension multimodale de la couleur. Ainsi, nous choisissons d'utiliser le concept de couleur pour faire référence à un objet dans la description textuelle. De plus, dans le jeu de données constitué, quand il y a deux objets dans l'image, leur couleur est donc différente.

2.1 Relations spatiales

Nous créons plusieurs tâches de sondage pour tester la compréhension spatiale de ces modèles à différents niveaux d'ancrage du texte dans l'image. Plusieurs relations spatiales sont considérées :

— La position d'un objet dans le cadre de référence de l'image ("gauche/droite" et "devant/derrière")
— La position d'un objet relativement à un autre objet dans le cadre de référence de l'image ("gauche/droite" et "devant/derrière")
— La distance distance entre deux objets dans le cadre de référence de l'image ("loin"/"proche")

La position des objets est annotée précisément en utilisant les coordonnées des pixels, afin de pouvoir sélectionner les images et générer des descriptions selon un modèle. Quand nous considérons la position d'un objet dans le cadre de référence de l'image, des limites sont définies pour marquer si un objet se trouve à gauche, à droite, devant ou derrière. Une marge est utilisée pour éliminer les situations trop ambiguës, avec des objets trop proche du centre. Le même type de marge est utilisée pour considérer la position relative d'un objet par rapport à l'autre, ainsi que la distance entre deux objets, quand la distinction est difficile à faire.

2.2 Tâches de sondages

Chaque image du jeu de données est associée à une description textuelle en anglais qui est créée automatiquement suivant un modèle de syntaxe. Ces descriptions varient en fonction des tâches considérées, afin d'évaluer plus précisément la compréhension des concepts de position et l'ancrage

des informations visuelles dans le texte. Ainsi, les tâches de sondage suivantes sont réalisées :

— *Compréhension monomodale de la position d'un objet dans l'image* : Ce sont deux tâches de classification binaire de la position d'un objet unique dans le cadre de référence de l'image. Les classes considérées sont "gauche"/"droite" ou "devant"/"derrière". L'utilisation d'informations visuelles est suffisante pour réaliser cette tâche.

— *Compréhension monomodale de la distance entre deux objets* : C'est une tâche de classification binaire de la distance entre deux objets dans l'image. Les classes considérées sont "loin"/"proche". L'utilisation d'informations visuelles est suffisante pour réaliser cette tâche.

— *Compréhension multimodale de la position d'un objet dans l'image* : Ce sont des tâches de classification binaire qui évaluent la position d'un objet dans le cadre de référence de l'image, pour des images à deux objets. de la position d'un objet référencé dans la légende dans des images à deux objets. La description textuelle consiste à porter référence à l'objet par sa couleur, et le modèle doit évaluer sa position. Le modèle doit utiliser les informations textuelles pour savoir quel est l'objet considéré, et les informations visuelles pour déterminer sa position.

— *Compréhension multimodale de la position relative d'un objet par rapport à un autre* : Ce sont deux tâches de classification binaire permettant d'évaluer si une légende décrit correctement la position relative de deux objets. L'ancrage dans l'image du texte est nécessaire, car le modèle doit utiliser les informations visuelles pour vérifier si la description textuelle de la position est vraie.

— *Compréhension multimodale de la distance entre deux objets* : C'est une tâche de classification binaire permettant d'évaluer si une légende décrit correctement la distance de deux objets. L'ancrage dans l'image du texte est nécessaire, car le modèle doit utiliser les informations visuelles pour vérifier si la description textuelle de la distance est vraie.

Les expériences sont réalisées en utilisant un modèle de sondage linéaire. Les résultats sont une moyenne de 5 essais avec des graines différentes.

3 Résultats

3.1 Évaluation de la position d'un objet dans le cadre de référence de l'image

Nous observons que les modèles ont une bonne compréhension monomodale de la position d'un objet dans le cadre de référence de l'image. Leur performance est légèrement inférieure à celle de modèles monomodaux VIT (Dosovitskiy *et al.*, 2020) ou ResNet (He *et al.*, 2016), comme le montrent les résultats du tableau 2. Nous pouvons voir à travers ces résultats que tous les modèles parviennent à extraire les informations relatives à la position de l'image. LXMERT a de meilleurs performances que les autres modèles.

Tâche	ViLT	UNITER	LXMERT	VIT	ResNet
Droite/gauche	83,94	93,21	96,76	90,66	99,4
Devant/derrière	97,04	92,64	98,61	98,47	99,31

TABLE 2 – Évaluation de la compréhension monomodale de la position d'un objet dans le cadre de référence de l'image (Accuracy).

Les modèles Vision-Langage ont une moins bonne performance dans les tâches multimodales de

compréhension de la position absolue, mais elle reste supérieure aux références monomodales (Tableau 3). UNITER a de moins bonnes performances que les autres modèles. [1]

Tâche	ViLT	UNITER	LXMERT	VIT	ResNet
Droite/gauche	65,92	56,09	73,12	52,8	50,44
Devant/derrière	92,23	82,28	89,31	80,22	75,86

TABLE 3 – Evaluation de la compréhension multimodale de la position d'un objet dans le cadre de référence de l'image

3.2 Évaluation de la position d'un objet relativement à un autre objet

La tache de compréhension multimodale de la position relative d'un objet montre les faiblesses de l'ancrage multimodal des modèles Vision-Langage. En effet, le tableau 4 montre que les résultats de ces modèles ne sont pas supérieurs aux références monomodales. Seul LXMERT obtient des résultats significativement supérieurs, et ce, seulement pour la tâche de classification entre "gauche" et "droite".

Tâche	ViLT	UNITER	LXMERT	VIT	ResNet
Gauche/droite	46,48	49,83	71,14	49,83	50,74
Devant/derrière	55,75	50,04	50,67	52,58	48,67

TABLE 4 – Évaluation de la compréhension multimodale de la position relative d'un objet par rapport à l'autre (Accuracy).

3.3 Évaluation de la distance entre deux objets

Nous comparons les résultats d'évaluation monomodale et multimodale de la distance entre deux objets dans le tableau 5. Ces résultats montrent que les modèles Vision-Langage parviennent à extraire les informations visuelles nécessaires pour évaluer la distance entre deux objets, même si les résultats obtenus sont inférieurs aux références visuelles monomodales VIT et ResNet. Cependant, l'évaluation multimodale montre qu'il n'y a pas d'ancrage des mots relatifs à la distance dans l'image.

Model	ViLT	UNITER	LXMERT	ResNet	VIT
Évaluation monomodale	66.44	70.10	74.23	92.79	66.92
Évaluation multimodale	52.50	50.77	51.44	52.60	44.42

TABLE 5 – Évaluations de monomodale and multimodale de la distance entre deux objets. Pour la première tâche, le texte n'utilise pas de mot relatifs à la distance pour différencier, seuls l'image peut fournir les informations. La deuxième tâche consiste à évaluer si une légende (par exemple, "l'objet rouge est à côté de l'objet bleu") est vraie (comparée à "l'objet rouge est loin de l'objet bleu") (Accuracy).

1. Les données pour la classification "devant"/"derrière" sont déséquilibrées, ce qui explique que VIT et ResNet aient des performances supérieures à 50%.

4 Analyse

Ces évaluations nous permettent de montrer que les modèles parviennent bien à extraire les informations relatives à la position d'un objet dans le cadre de référence de l'image, mais qu'ils ont plus de mal à la combiner à l'information textuelle pour en extraire une information multimodale. Notamment, les tâches évaluant la compréhension multimodale des positions relatives entre deux objets sont plus complexes que la position d'un objet dans le cadre de l'image.

De plus, on remarque que LXMERT a une meilleure compréhension multimodale des concepts spatiaux. Or, LXMERT, contrairement à UNITER et ViLT, utilise "Visual Question Answering" (VQA) (Zhang *et al.*, 2016) comme tâche de pré-entraînement, en combinant plusieurs jeux de données. Une analyse de ces données montre qu'elles contiennent de nombreuses de questions en rapport avec la position, notamment sur les concepts de "gauche" et "droite", que LXMERT parvient relativement bien à ancrer dans l'image.

Ainsi, si ces modèles peuvent extraire les informations visuelles relatives à la position d'un objet dans l'image relativement facilement, et à les utiliser dans un contexte multimodal, ils ont plus de mal à extraire les informations liées à la position relative des objets et à ancrer dans l'image les mots liés à la position. Les différences de performances des modèles semblent montrer que cela n'est pas du à l'architecture des modèles, car les informations visuelles relatives à la position sont bien extraites par les modèles, mais aux tâches multimodales de pré-entraînement utilisées. Une tâche multimodale avec des annotations précises nécessitant un raisonnement lié à la position comme VQA utilisée par LXMERT semble nécessaire pour ancrer les mots relatifs à la position dans l'image.

5 Conclusion

Nous avons étudié pourquoi les modèles Vision-Langage ont une faible compréhension multimodale de la position. En sondant les modèles, nous observons que ces faiblesses ne sont pas principalement dues à l'architecture des modèles, aux représentations visuelles ou à une mauvaise représentation de la position des objets, mais aux tâches et données de pré-entraînement. En effet, l'utilisation de tâches, telles que VQA, nécessitant un raisonnement spatial semble indispensable pour ancrer les mots relatifs à la position dans l'image. De futurs travaux pourront donc étudier la construction de telles tâches afin d'améliorer la compréhension multimodale de ces concepts.

Références

CHEN Y.-C., LI L., YU L., EL KHOLY A., AHMED F., GAN Z., CHENG Y. & LIU J. (2020). Uniter : Universal image-text representation learning. In *European conference on computer vision*, p. 104–120 : Springer.

DOSOVITSKIY A., BEYER L., KOLESNIKOV A., WEISSENBORN D., ZHAI X., UNTERTHINER T., DEHGHANI M., MINDERER M., HEIGOLD G., GELLY S. *et al.* (2020). An image is worth 16x16 words : Transformers for image recognition at scale. *arXiv preprint arXiv :2010.11929*.

GOYAL Y., KHOT T., SUMMERS-STAY D., BATRA D. & PARIKH D. (2017). Making the V in VQA matter : Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

HE K., ZHANG X., REN S. & SUN J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 770–778.

JOHNSON J., HARIHARAN B., VAN DER MAATEN L., FEI-FEI L., LAWRENCE ZITNICK C. & GIRSHICK R. (2017). Clevr : A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 2901–2910.

KIM W., SON B. & KIM I. (2021). Vilt : Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, p. 5583–5594 : PMLR.

RÖSCH P. J. & LIBOVICKỲ J. (2022). Probing the role of positional information in vision-language models. In *Findings of the Association for Computational Linguistics : NAACL 2022*, p. 1031–1041.

SALIN E., FARAH B., AYACHE S. & FAVRE B. (2022). Are vision-language transformers learning multimodal representations ? a probing perspective. In *AAAI 2022*.

SUHR A., ZHOU S., ZHANG I., BAI H. & ARTZI Y. (2018). A corpus for reasoning about natural language grounded in photographs. *CoRR*, **abs/1811.00491**.

TAN H. & BANSAL M. (2019). Lxmert : Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv :1908.07490*.

ZHANG P., GOYAL Y., SUMMERS-STAY D., BATRA D. & PARIKH D. (2016). Yin and Yang : Balancing and answering binary visual questions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

The application of natural language processing (NLP) tools in relation to selected Mongolic languages: review of the current literature, available NLP tools and outlooks for the future

Joanna Dolińska[1]
(1) University of Warsaw, ul. Krakowskie Przedmieście 26/28, 00-927 Warsaw, Poland
j.dolinska-streltsov@uw.edu.pl

RESUME

Cet article aborde le problème des langues mongoles sélectionnées : le Khamnigan Mongol, le Oirat Mongol et le Dagur, parlées en Russie, en Mongolie et en Chine. L'objectif de cet article est de répondre aux questions suivantes : Quel est l'état sociolinguistique actuel de ces langues et quels outils de traitement automatique de la langue naturelle (TALN) ont été appliqués à quelles langues mongoles jusqu'à présent ? L'article présenté démontrera que le développement d'outils de TALN pour la famille des langues mongoles n'a commencé que récemment et concerne principalement les tâches de recherche d'information (RI) et les sujets qui y sont liés, tels que l'utilisation de la recherche du radical et des listes d'arrêt dans la RI, le système de recherche par mots-clés pour localiser des mots dans des images de documents historiques mongols, les unités de recherche basées sur les n-grammes et les outils de reconnaissance vocale consacrés au phénomène mongol de l'harmonie des voyelles.

ABSTRACT

The application of natural language processing (NLP) tools in relation to selected Mongolic languages: review of the current literature, available NLP tools and outlooks for the future

This paper tackles the problem of the selected Mongolic languages: Khamnigan Mongol, Oirat Mongol and Dagur spoken in Russia, Mongolia and China. The aim of this presentation is to answer the following questions: What is the current sociolinguistic condition of these languages and which NLP tools have been applied to which Mongolic languages so far? The presented paper will demonstrate that the development of the NLP tools for the family of Mongolic languages started only recently and refers mostly to the information retrieval (IR) tasks and topics which are related to it, such as applying stemming and stoplist in IR, keyword retrieval system for locating words in historical Mongolian document images, n-gram-based retrieval units and speech recognition tools devoted to the Mongolian phenomenon of vowel harmony.

MOTS-CLES : Langues indigènes, Khamnigan Mongol, Oirat Mongol, Daur, Langues en danger, TALN

KEYWORDS: Indigenous languages, Khamnigan Mongol, Oirat Mongol, Dagur, Endangered languages, NLP

1 Introduction

Safeguarding cultural and linguistic diversity belongs to the Sustainable Development Goals of the United Nations. One may argue that the birth, development and death of languages constitute

Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain *(LIFT) et* Traitement Automatique des Langues *(TAL),*
Marseille, France, 14 au 15 novembre 2022 , pages 188–197,

natural stages of a language's life. Nevertheless, some languages become extinct much faster than the others due to economic, historical and technological circumstances. Modern technology, rising social awareness of the importance of linguistic variety, as well as the growing sentiment for indigenous cultures, which develops in a parallel way to globalization, create perfect conditions for the documentation and digitization of endangered and vulnerable languages.

Endangered and vulnerable languages can certainly be found also among the Mongolic languages (Yu, 2011). Together with the Tungusic, Turkic, Koreanic and Japonic families of languages they constitute the group of languages called "Transeurasian" (Robbeets et al., 2021) which has been traditionally known under the name "Altaic" (Janhunen, 2005). There are currently 98 Transeurasian languages in the world spread across Northeast Asia, Siberia, Central Asia, Anatolia and some parts of Europe. The Mongolic language distribution stretches from the Caspian Sea region in the west to Manchuria in the east. Vertically, it covers the area from the Baikal region up to the northern Afghanistan and the Gansu/Qinghai region of China in the south. For the purpose of this paper three endangered Mongolic languages have been selected for description: Khamnigan Mongol, Oirat Mongol and Dagur.

To start with, it is worth clarifying the terms „Mongolian" and „Mongolic". The term "Mongolic" refers to the family of languages which disintegrated into various languages around the 13th century, when Chinggis Khan started his military conquests, whereby "Mongolian" means the language spoken in Mongolia and in the region of Inner Mongolia, China. In 2007 there were approximately 9,6 million Mongolic people in the world. Nevertheless, only ca. 50% of individuals who consider themselves to be of Mongolic origin, can actually speak their language (Rybatzki, 2020: 24). There are several classifications of Mongolic languages, based on the geographical and linguistic criteria. One of the most recent classifications of the Mongolic languages has been proposed by J. Janhunen (2012: 3). They are divided into four branches: I The Dagur branch in the historical Manchuria (Dagur language), II The Common Mongolic branch in the present day Mongolia, parts of Siberia, Manchuria, Ordos and Dzungaria (Khalkha Mongolian and its dialects), III The Shirongolic branch in the Gansu and Qinghai Provinces of China (Shira Yughur, Mongghul, Mongghuor, Mangghuer, Bonan, Kangjia and Santa), IV The Moghol branch in Afghanistan (Moghol language with several local varieties, whereby this branch has almost certainly died out). Another common classification of Mongolic languages has a geographic character and it stipulates the existence of the so called core Mongolic languages (Buryat, Khamnigan Mongol, Khalkha, Ordos, Kalmyk and Oirat) and the peripheral Mongolic languages (Moghol, Dagur, Shira Yughur, Santa Mongghul and Baoan).

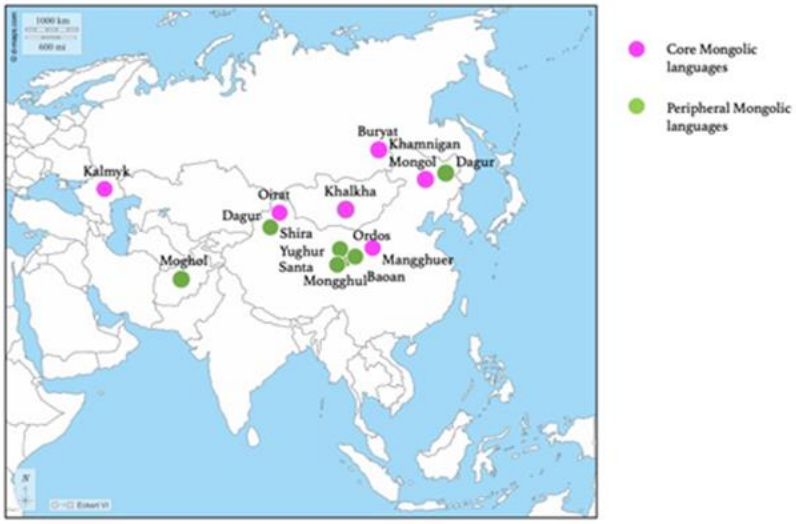


Figure 1: Core and peripheral Mongolic languages. (Localization of the languages on the map has been carried out by the author of this article. The map comes from the free map stock d-maps.com.)

Nevertheless, the division between the "core" and "peripheral" languages is not flawless, as the "peripheral" Mongolic languages are often understood as "all languages except Mongol proper, Buryat and Oirat-Kalmyk" or „mostly non-written, politically unimportant and insufficiently documented languages" (Nugteren, 2020).

1.1 Description of the Khamnigan Mongol, Oirat Mongol and Dagur

Khamnigan Mongol people, also called "Horse Tungus", inhabit the area of Transbaikalia, even though they originally came from the northeast area of Mongolia where Chinggis Khan was born (Janhunen, 2005: 82). Some of them migrated to the northern part of China in the early 20th century. From the 17th century onwards, the Khamnigan Mongol people have been under the Russian influence. Furthermore, they form an ethnic group where Mongolic and Tungusic influences are blended. In China, the Khamnigan community amounts to 2000 persons (but this number includes also Tungusic speaking Khamnigan people). Khamnigan language was recognized as a separate language only very recently, namely in the 20th century. Historically (Janhunen, 2005), this language has been described by Jamtsarano & Damdinov (1982), L. Mishig (1961), B. Rinchen (1969) and J. Janhunen (1992; 1996).

The Oirat group of languages is also known under the term "western Mongolic languages." Oirat languages are spoken in the Republic of Kalmykia in Russia (Kalmyk Oirat), in the Xinjiang province in China (Xinjiang Oirat), and in Mongolia (the so called "Altai Oirat"). The main Oirat Mongolian communities in Mongolia are settled in the Uws and in the Khowd provinces. Some Oirats can also be found in the westernmost parts of Mongolia and in the Arkhangai province (Birtalan, 2020: 351). Nevertheless, it is assumed that these communities stopped speaking Oirat to the advantage of Khalkha Mongolian, although they preserved their Oirat identity. According to the last comprehensive census in Mongolia, 9% of its total population considered themselves to be Oirat Mongolian (Birtalan, 2020: 351). According to J. Janhunen (2005), the Oirat languages have been historically researched, among others, by Tsoloo (1965), Wandui (1965), I. Ya. Zlatkin (1964) and H. Okada (1987).

Dagur language is spoken in the easternmost region of all Mongolic languages (Yamada, 2020: 321). It can be estimated that there are around 100 000-110 000 Dagur speakers worldwide. The so called Dagur Nonni speakers, also known as Butha Dagur, are present in Morin Dawa Daur Autonomous Banner in Hulun Buir League, the Nenjiang and Qiqihar regions in Inner Mongolia, where ca. 100 000 persons spoke this language in 2003 (Tsumagari, 2003). The Butha Dagur dialect is perceived as the standard dialect of Dagur (Yamada, 2020: 321).The number of the so called Dagur (Hailar) speakers, who are present in the Ewenki Autonomous Banner in Hulun Buir League south of the city of Hailar, amounted to 5000 in 2003 (Tsumagari, 2003). The so called Turkestan Dagur or Xinjiang Dagur speakers live in the Xinjiang Uygur Autonomous Region with the approximate number of 4000 speakers (Tsumagari, 2003). Historically, the Dagur population inhabited the Middle Amur region. The Dagur language has no official literary language (Janhunen, 2005: 129), yet historically certain Dagur poets used Manchu script to write down their works (Yamada, 2020: 321). The Dagur language, its grammar and lexicon have been historically researched (Janhunen, 2005) among others by B. H. Todaeva (1986; 1997), N. Poppe (1930; 1934-

1935; 1964), Erhimbayar and Enhebatu (1988), as well as by S. Kałużyński (1969-1970).

The Khamnigan Mongol has a very limited number of speakers and it is definitely in need of protection. This low-resource language is spoken at the same time in two countries: Russia and China. Its speakers are usually at least bilingual and they speak a Tungusic language as well. When it comes to Oirat Mongol, the Oirat dialects have a long history and the Oirat ethnic group has been attested in the earliest known Mongolic epic work "The Secret History of the Mongols" from the 13th century. Oirat languages have been relatively well researched in Kalmykia and in Xinjiang, but it seems that in Mongolia there is still a need for further research, especially the modern sociolinguistic background of the Oirat Mongol communities. The reasons for paying special attention to these selected Mongolic languages are as following. The Dagur language has been selected for the analysis because it is endangered, yet there is a considerable number of its speakers to consult. The fact of missing the official literary language makes it interesting to focus on the oral literature and it points to the special potential of analyzing audio data in this language.

2 Modern research focused on the endangered Mongolic languages

In the years 2003-2009 fieldwork studies on 30 endangered Mongolic languages and dialects was carried out by the "Altaic Society of Korea, Researchers on Endangered Altaic Languages (ASK REAL)". Within this project, the research on the Qiqihar Dagur language took place in the Qiqihar City, Heilongjiang Province in China in 2003. The so called Xinjiang Dagur was researched in Tacheng City, Xinjiang Uyghur Autonomous Region in 2004. One of the conclusions of this project was that "all Mongolic languages and dialects, with the exception of Khalkha Mongolian, are facing a danger of disappearing under the increasing influence of Khalkha Mongolian, Tsakhar Mongolian, Chinese, Russian, Kazakh, or even Tibetan" (Yu, 2011: 276). Another regional perspective on the Mongolic languages was taken by Sarala Puthuval (2017) who analyzed the gradual stages of language shift in the 20th century Inner Mongolia. The quantitative research was carried out by this author with the help of a questionnaire which was applied in over 600 interviews. The main conclusion of this research is that the Mongolian-Chinese bilingualism shift into Chinese monolingualism in the 20th century. Another project, which describes, among others, the endangered Mongolic languages, is the "Endangered Languages Project" (ELP). It serves as an "online resource for samples and research on endangered languages" and at the same time it offers a platform for advice and best practices when working with linguistic diversity. When it comes to Kalmyk Oirat, which belongs to the same group of Oirat languages as Mongolian Oirat language, the cultural heritage of its speakers has been the subject of the "The Kalmyk Cultural Heritage Documentation Project" carried out at the University of Cambridge in the years 2014-2019 and it has been concentrated on the topics of Kalmyk kinship, intangible cultural heritage, material culture, economy, religion and traditional medicine.

3 The application of the NLP tools to the modern Mongolic languages – a review

The analysis of the scientific publications concerning the application of the NLP tools to the Mongolic languages points to two striking facts. First of all, the research on using the NLP tools in the area of Mongolic languages started only at the end of the first decade of the 21st century. G. Gao, W. Jin, F. Long & H. Hou stated in their article (2008): "By now, there is no functional information retrieval (IR) system available because the computational processing of Mongolian started very late". The problems which were imminent at the beginning of the NLP research on the Mongolic languages referred to the limited sources of digital Mongolian texts which could be used in the

information retrieval research. The authors of the above-mentioned article opted for using the digital resources of the newspapers Inner Mongolia Ordos Daily and Inner Mongolia Daily. Even in this very early stage of research on the IR in Mongolian it has been stated that the particular features of the Mongolian language require specific approach, as Mongolian is an agglutinative language with a vocal harmony. In Mongolian, words are created with the help of affixes that are "glued" to the base of the word. Sometimes a word can consist of a base and even three or more suffixes attached to it. For the purposes of the information retrieval, the stemming and stoplist methods were applied. The conclusion of this research article was that word stemming in Mongolian IR is important, because it can improve retrieval effectiveness. However, using stoplist can only slightly improve retrieval effectiveness, yet it can reduce the index in a significant way (Gao et al. 2008).

Another NLP research article (Yue, Gao and Min, 2015) is devoted to the search for the selection method of retrieval units in the Mongolian IR system, whereby the agglutinative character of the Mongolian language has been taken into special consideration. The authors of the article conclude that the retrieval units for Mongolian can be divided into stem + 1 affix, stem + 2 affixes and stem + 3 affixes. The IR models which have been applied in this research were TF-IDF Model, vector space model and the language model to explore the retrieval effect on root + 1 affix, root + 2 affixes and root + 3 affixes. The authors found out that best performance for IR in Mongolian is achieved by root + 2 affixes under the Lemur language model. Given the fact that the retrieval efficiency greatly depends on the choice of the proper retrieval unit, J. Yue, G. Gao and L. Min (2015) also carried out research on the n-gram based IR units in Mongolian, where for n-gram form, n ranges from 2 to 5, KL-Divergence, as well as two smoothing algorithms (Good-Turing Smooth and JM Smooth) were applied. The authors concluded that both smoothing algorithms are suitable for IR in Mongolian, whereby the best performance is reached in the form of n-gram (n=4).

When it comes to the traditional Mongolian script, the NLP tools have also been applied in the keyword retrieval system for historical Mongolian document images by H. Wei & G. Gao (2014). The above mentioned article shows a conscious approach to the use of the language processing technology in order to preserve the cultural heritage of the Mongolian people. The authors of this article analyzed the Mongolian Kanjur, which was originally translated from Tibetan. This work can be perceived as a "Mongolian encyclopedia" (Wei & Gao, 2014), whereby it includes information on the religion, history and literature of the Mongolian people and was made by woodblock printing. The authors of the article applied the word spotting technology by converting document images into a collection of word images with the help of word segmentation. Furthermore, a number of profile-based features were extracted to display word images. The authors developed a system which supports image-to-image matching through the calculation of similarities between a query word image and each word image in the researched text. The result of such a matching was a descending order of the similarities, whereby the query word image could be produced by the synthesis of a sequence of retrieved glyphs. When working with the keyword retrieval system for historical Mongolian document images the authors also noticed that the Mongolian-specific features need to be taken into consideration, for example the nature of the affixes glued to the word base in Mongolian. Some affixes are applied to create new words, but a considerable group of affixes is applied for inflectional purposes. The authors of the article suggest that refining the division between these affixes in the future research by dropping the inflectional affixes could increase the efficiency of their keyword retrieval system.

Another issue that has bothered the researchers working on applying the NLP tools to Mongolian refers to the character of the traditional Mongolian script which is still in use in Inner Mongolia and is regaining its popularity in Mongolia as well, where the Cyrillic script prevails. The traditional

Mongolian script has been placed in Unicode at the range of 1800 — 18AF and it is written from top to down from left to right, which is why it may cause display failures in general operating systems (Yiru, 2016). Another interesting feature which might complicate the processing of the traditional Mongolian script is that the shape of the most letters differs depending on whether they are on the onset, in the middle or on the outset of the word, which resembles the feature, for example, of the Arabic script.

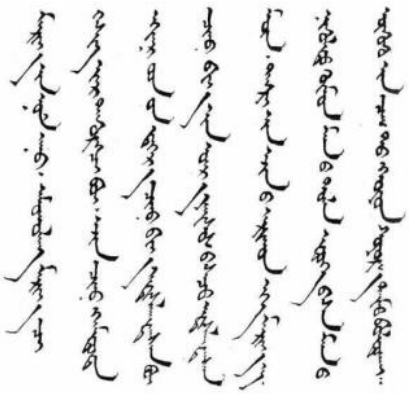


Figure 2: An example of the traditional, written Mongolian script. Source: Source: https://mininmongolia.wordpress.com/tag/traditional-mongolian-script/

Another striking observation is that the NLP tools have been so far applied only to those Mongolic languages, which have the highest number of speakers, i.e. to Mongolian spoken in Inner Mongolia, China and to Khalkha Mongolian spoken in the country of Mongolia. So far, the review of the scientific literature shows the other Mongolic languages presented in the Introduction part of this article have not been the subject of the NLP tasks.

Although to a considerably smaller extent than in the case of the Classical Mongolian and the Mongolian language spoken in Inner Mongolia, China, there are several publications devoted to the Khalkha Mongolian language spoken in the country of Mongolia. An example of such a work is the article of U. Banjaree, I. Dutta and Irfan S. devoted to the topic of vocal harmony in Mongolian, which is yet another topic very specific to the Khalkha Mongolian language (2017). This feature of the language refers to the regularity of vowels that appear in the Khalkha Mongolian language. To sum up this feature in a brief way, if a given type of a vowel is chosen at the onset of a given word, the suffixes following the base of the word will also include the same type of a vowel. This feature is common to all Mongolic languages, only the frequency of its appearance varies from language to language.

Furthermore, C. Hansakunbuntheung et al. presented in their paper “Mongolian speech corpus for text-to-speech development” (2011) their first attempt to develop a Mongolian speech corpus, the aim of which is to support the development of a Hidden-Markov-Model-based (HHM-based) text to speech (TTS) system for a screen reader. The aim of this research was to improve the quality of life of Khalkha Mongolian speakers with the vision impairment. The high value of this afore-mentioned article lies, among others, in the fact that it lists the already present Khalkha Mongolian corpora, which include a large Mongolian text corpus that encompasses 5 million words and which has been collected at the National University of Mongolia. This corpus is semi-automatically tagged by a spell checker and parsing tools. The authors also mention A Multi-dialectal speech corpus of Mongolia (MDSCM) which preserves and divides the Mongolian text and speech dialects in three groups, namely Khalkha, Chakhar and Oirat groups of dialects. There authors also mention the third

speech corpus, which is devoted to the automatic speech recognition. It consists of a list of 10 words which have been repeatedly recorded (10 times each word) by 10 speakers. Therefore, this example constitutes a rather small corpus. The authors of the article conclude that their speech corpus was designed to cover all Mongolian phones, while providing phonetic transcription with stress marking and the context information for further modeling of speech acoustics. One of the remaining challenges mentioned by the authors is the grapheme to phoneme (G2P) conversion of the foreign words in the Khalkha Mongolian language.

When it comes to the Khalkha Mongolian language, not only the topic of corpora creation and the development of a TTS system has been in the focus of the scientific research. W. Wang, F. Bao and G. Gao (2015) published an article on the Mongolian Named Entity Recognition (NER) in which they stipulate that their research can also be applied to other NER systems devoted to different agglutinative languages.

Finally, it is worth mentioning an article devoted to the application of the NLP tools to the Buryat language, which is spoken mainly in the region of the Republic of Buryatia in the Russian Federation. The publication of O. Rinchinov (2019) is devoted to the preparation of textual data for the diachronic corpus of the Buryat language on the bases of Buryat chronicles. These chronicles have been written down in the Mongolian script – the same which is used for writing down the Mongolian texts in Inner Mongolia. The main point of discussion in this article are the main parameters of the structural markup of textual data which are based on punctuation markers that have been introduced in the process of the Latin transliteration. This article provides an evidence that the NLP research on the Mongolic languages is expanding from the tight circle of high-resource Khalkha Mongolian and Mongolian spoken in Inner Mongolia, China, to less distributed languages, such as Buryat. Nevertheless, the Buryat language is still spoken by a considerable number of speakers. According to E. Skribnik (2005), the census from 1989 showed that there were around 363 000 Buryat speakers in the Russian Federation at that time.

4 Conclusions

The analysis of the scientific publications concerning the application of the NLP tools to the Mongolic languages highlights two interesting facts. First of all, the research in this field started relatively late. It was inspired by the research methods introduced in the study of Western and Arabic languages (Yue, Gao and Min, 2015). Secondly, it has been primarily focused on high-resource languages (Mongolian spoken in Inner Mongolia, China and Khalkha Mongolian spoken in Mongolia, as well as Classical Mongolian). Publications on less distributed Mongolic languages, such as Buryat, are scarce. The most important features of the Mongolic languages highlighted in the NLP research so far refer to the agglutinative nature of these languages, the vowel harmony and the features of the traditional written Mongolian script (written from top to down, from left to right).

The endangered Mongolic languages presented at the beginning of this article (Khamnigan Mongol, Oirat Mongol and Dagur) need to be protected. All of them are low-resource languages and represent interesting features that reveal the long history of various speech communities that have been crossing the borders of Russia, Mongolia and China over the centuries. Khamnigan Mongol is spoken both in China and Russia, whereby its speakers are at least bilingual, because they also speak a Tungusic language. Oirat Mongol belongs to the old group of Oirat family that has been attested in the Mongolian epic work from the 13th century “The Secret History of the Mongols”. The Dagur language does not have a literary version and is the easternmost Mongolic language. This article advocates for the use of the NLP tools in order to preserve these and endangered Mongolic languages. The author of this article also envisages taking steps in the future to start an

NLP research on these languages.

Last but not least, it is worth mentioning the arguments which support the thesis that the application of the NLP tools can support the survival and maintenance of the endangered languages. First of all, the NLP tools make the endangered languages more accessible to a broader audience. Moreover, they can also help reignite the interest of the descendants of given language communities to refresh their knowledge of their native language. Furthermore, through the NLP tools not only the lexicon, semantics and grammar of a given language are preserved for the future generations, but also the phonological and phonetic features of this language. In addition, if a speaker community does not have a written language (for example Dagur), the NLP tools offer the possibility to collect and analyze the oral literature of this community. Last but not least, if the indigenous languages are included in the worldwide research on the application of the NLP tools, it may increase the social status of their speakers in a long term perspective.

Références

BANERJEE, U., DUTTA, I. & IRFAN S. (2017). Coarticulatory propensity in Khalkha Mongolian. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, p. 356–361. NLP Association of India.

BIRTALAN, A. (2020). Oirat and Kalmyk, the Western Mongolic languages. In M. ROBBEETS & A. SAVELYEV, Éds., *The Oxford Guide to the Transeurasian Languages*, p. 350-369. DOI : https://doi.org/10.1093/oso/9780198804628.001.0001.

ENDANGERED LANGUAGES PROJECT. URL : www.endangeredlanguages.com

ERHIMBAYAR [EERHENBAYAER] & ENHEBATU. (1988). *Dawoeryu Duben*, Hohhot. Neimenggu Jiaoyu Chubanshe.

GAO, G., JIN, W., LONG, F., & HOU, H. (2008). A First Investigation on Mongolian Information Retrieval. In *EVIA@ NTCIR*. URL: (EVIA2008).

HANSAKUNBUNTHEUNG, CH., THANGTHAI, A., THATPHITHAKKUL, N. & CHAGNAA, A. (2011). Mongolian speech corpus for text-to-speech development. In *International Conference on Speech Database and Assessments (Oriental COCOSDA)*, p. 130-135. DOI : 10.1109/ICSDA.2011.6085994.

JAMTSARANO, TS. J. & DAMDINOV, D. G. (1982). *Uligery ononskix xamnigan*, Novosibirsk. Nauka.

JANHUNEN, J. (1992). On the Position of Khamnigan Mongol. *Journal de la Société Finno-Ougrienne* 84, 115-143.

JANHUNEN, J. (1996). Mongolic Languages as Idioms of Intercultural Communication in Northern Manchuria. In S. A. WURM, P. MÜHLHÄUSLER & D. T. TRYON, Éds., *Atlas of Languages of Intercultural Communication in the Pacific, Asia, and the Americas*, II.2, p. 827–34.

JANHUNEN, J. (2005). *The Mongolic languages*, London. Routledge.

JANHUNEN, J. (2012). *Mongolian*, Amsterdam/Philadelphia. John Benjamins Publishing Company.

KALMYK CULTURAL HERITAGE DOCUMENTATION PROJECT. URL : https://www.kalmykheritage.socanth.cam.ac.uk/en/index.php?language=en

KAŁUŻYŃSKI, S. (1969–70). Dagurishes Wörterverzeichnis nach F. V. Muromskis handschriftlichenSprachaufzeichnungen. *Rocznik Orientalistyczny* 33/1, 103–44, 33/2, 109–43.

MISHIG, L. (1961). Mongol Ard Ulsïn zarim nutgiin xamnigan ayalguug surwaljilsan ny. In *Olonulsïn mongol xel bichgiin erdemtnii anxdugaar ix xural 1*, p. 183–203, Ulaanbaatar.

NUGTEREN, H. (2020). The classification of Mongolic languages. In M. ROBBEETS & A. SAVELYEV, Éds., *The Oxford Guide to the Transeurasian Languages*, p. 92-104. Oxford University Press. DOI : https://doi.org/10.1093/oso/9780198804628.001.0001.

OKADA, H. (1987). Origin of Dörben Oyirad. *Ural-Altaische Jahrbücher*, 7, 181–211.

OLKO, J., & SALLABANK, J. (2021). *Revitalizing Endangered Languages: A Practical Guide*, Cambridge. Cambridge University Press. DOI :10.1017/9781108641142.
POPPE, N. (1930). *Dagurskoe narechie [Dagur]*. Izdatel'stvo Akademii Nauk SSSR.
POPPE, N. (1934–5). Über die Sprache der Daguren. *Asia Major*, 10, 1–32, 183–220.
PUTHUVAL S. (2017). Stages of language shift in twentieth-century Inner Mongolia, China. *Proceedings of the Linguistic Society of America*, 2, 28, 1-14. DOI : https://doi.org/10.3765/plsa.v2i0.4083
POPPE, N. (1964). Die dagurische Sprache. *Mongolistik, Handbuch der Orientalistik I*, V, 2, 137–42.
RINCHEN, B. (1969). *Mongol Ard Ulsïn xamnigan ayalguu*, Ulaanbaatar. BNMAU Shinjlex Uxaanï Akademiin Xewlel.
RINCHINOV, O. (2019). Structural Markup of the Mongolian-script Buryat chronicles for the diachronic corpus of Buryat Language. *Culture of Central Asia: written resources*, 12, 106-117. DOI : 10.30792/2304-1838-2019-106-117.
ROBBEETS, M., BOUCKAERT, R., CONTE, M., SAVELYEV, A., LI, T., AN, D.-I., SHINODA, K., CUI, Y., KAWASHIMA, T., KIM, G. UCHIYAMA, J., DOLINSKA, J., OSKOLSKAYA, S., YAMANO, K.-Y., SEGUCHI, N., TOMITA, H., TAKAMIYA, H., KANZAWA-KIRIYAMA, H., OOTA,H., ISHIDA, H., KIMURA, R., SATO, T., KIM, J-H., BJØRN, R., DENG, B., RHEE, S., AHN, K.-D., GRUNTOV, I., MAZO, O., BENTLEY, J., FERNANDES, R., ROBERTS, P., BAUSCH, I., GILAIZEAU, L., YONEDA, M., KUGAI, M., BIANCO, R., ZHANG, F., HIMMEL, M., KRAUSE, J., HUDSON, M. & NING, C. (2021): Triangulation supports agricultural spread of the Transeurasian languages. *Nature*, 599, 616–621. DOI : https://doi.org/10.1038/s41586-021-04108-8.
RYBATZKI, V. (2020). The Altaic languages. Tungusic, Mongolic, Turkic. In M. ROBBEETS & A. SAVELYEV, Éds., *The Oxford Guide to the Transeurasian Languages*, p. 21-28. Oxford University Press. DOI : https://doi.org/10.1093/oso/9780198804628.001.0001.
SKRIBNIK, E. (2005). Buryat. In J. JANHUNEN, Éd., *The Mongolic Languages*, p. 102-128. Routledge.
TODAYEVA, B. KH. (1986). *Dagurskij jazyk* [Dagur]. Nauka.
TODAYEVA, B. KH. (1997). Dagurskii yazyk. In *Mongol'skie yazyki – Tunguso-man'chzhurskie yazyki – Yaponskii yazyk – Koreiskii yazyk* [Yazyki Mira], p. 51-60. Rossiiskaya Akademiya Nauk; Izdatel'stvo Indrik.
TSOLOO, J. (1965). *Zaxcinii aman ayalguu*, Ulaanbaatar. BNMAU-ïn ShUA Xel Zoxiolïn Xüreelen.
TSUMAGARI, T. (2003): Dagur. In J. JANHUNEN, Éd., *The Mongolic Languages*, p. 129-153. Routledge.
WANDUI, E. (1965). *Dörwöd aman ayalguu*, Ulaanbaatar. BNMAU-ïn Shinjlex Uxaanï Akademi.
WANG, W., BAO, F. & GAO, G. (2015). Mongolian named entity recognition using suffixes segmentation. In *2015 International Conference on Asian Language Processing*, IALP 2015, Suzhou, China, *October 24- 25*, p. 169-172. DOI : 10.1109/IALP.2015.7451558.
WEI, H. & GAO, G. (2014). A keyword retrieval system for historical Mongolian document images. *Document Analysis and Recognition. International Journal on Document Analysis and Recognition*, 17/1, 33-45. DOI : 10.1007/s10032-013-0203-6.
YAMADA, Y. (2020). Dagur. In M. ROBBEETS & A. SAVELYEV, Éds., *The Oxford Guide to the Transeurasian Languages*, p. 321-333. Oxford University Press. DOI : https://doi.org/10.1093/oso/9780198804628.001.0001.
YIRU (2016). Mongolia Language Resource Assessment. URL : (Mongolian-LRA)
YU, W. (2011). *A study of the Mongol Khamnigan spoken in Northeastern Mongolia*, Seoul. Seoul Natl. Univ. Pr.
YUE, J., GAO, G. & MIN, L. (2015). Study on Root + Affix Form-Based Mongolian Information

Retrieval Unit. *Proceedings of the 2015 3rd International Conference on Management Science, Education Technology, Arts, Social Science and Economics*. DOI：10.2991/msetasse-15.2015.155.
ZLATKIN, I. YA. (1964) *Istoriya Dzhungarskogo xanstva (1635–1758)*, Moskva. Nauka.

Vers la génération automatique de gloses pour la documentation automatique des langues

Shu Okabe[1] François Yvon[1]
(1) Université Paris-Saclay, CNRS, LISN, Bât. 508, Rue du Belvédère, F-91405 Orsay, France
shu.okabe@limsi.fr, francois.yvon@limsi.fr

RÉSUMÉ

Une étape du processus de la documentation d'une langue consiste à annoter des énoncés recueillis sur le terrain – après enregistrement et transcription phonétique – au niveau des morphèmes. Concrètement, pour chaque unité minimale segmentée dans la séquence d'entrée, il s'agit d'attacher soit une (plus rarement) plusieurs étiquettes morphosyntaxiques, soit une étiquette de concept, le plus souvent représenté par le mot anglais correspondant. Dans la perspective d'automatiser cette phase d'annotation, nous présentons les résultats d'une étude préliminaire où nous la considérons comme une tâche d'étiquetage de séquences, dont nous chercherons à estimer la difficulté, en la comparant à une tâche d'étiquetage morphosyntaxique standard. La question principale qui nous anime étant d'évaluer la faisabilité de cette annotation lorsque les données d'apprentissages sont très limitées.

ABSTRACT

Towards Automatic Gloss Generation for Computational Language Documentation.

One step of the language documentation process consists in annotating utterances collected on the field—once transcribed—at the morpheme level. For each minimal unit segmented in the input stream, the annotation process associates either one (or, in rare cases, several) morpho-syntactic tag(s) or one conceptual label represented by the corresponding English lemma. With the goal of automating this annotation task, we report the results of a preliminary study where this task is viewed as a sequence labelling task. Comparing the obtained results with a standard PoS task on the same data allows us to assess the difficulty of the process, especially in the context where the training resources are limited.

MOTS-CLÉS : génération de gloses interlinéaires, documentation automatique des langues.

KEYWORDS: interlinear gloss generation, computational language documentation.

1 Introduction

La documentation automatique des langues s'intéresse aux méthodes et outils destinés à assister les linguistes de terrain dans leurs tâches de collecte et d'annotation de données linguistiques, comme illustré Figure 1. Une fois transcrit phonétiquement, un énoncé (S) est segmenté en mots (séparés par les espaces) et morphèmes (séparés par des tirets). La glose interlinéaire (G) associe à chaque morphème une étiquette correspondant soit à son sens *lexical* exprimé par un concept de la langue cible (c'est le cas de *man* sur la figure), soit à sa fonction *grammaticale* (DEF dans l'exemple de la figure). Dans cet exemple, la dernière ligne (T) correspond à la traduction de la phrase dans une langue cible, ici l'anglais.

S	Phrase segmentée	bečedaw–ni	žek'u	razi	oq–n
G	Glose	wealthy–DEF	man	agree	become–PST.UNW
T	Traduction (EN)	*The wealthy man agreed.*			

FIGURE 1 – Strates d'annotation, extraites d'un corpus en langue tsez (Abdulaev & Abdulaev, 2010)

Notre objectif ultime est de produire automatiquement les gloses à partir de la phrase segmentée dans la langue source et de sa traduction en langue cible, deux ressources systématiquement disponibles. Les gloses constituent une strate d'annotation essentielle dans l'optique d'élaborer une grammaire ou un dictionnaire. Néanmoins, cette annotation est délicate et coûteuse à réaliser ; son automatisation, même partielle, permettrait d'accélérer cette étape. Le travail exploratoire présenté ici vise à évaluer la difficulté de cette tâche, par comparaison à la tâche d'étiquetage en partie du discours (PoS), dans un cadre de documentation des langues. Ce contexte implique des corpus de taille réduite, et des jeux d'étiquettes très grands, potentiellement « ouverts » (puisque les étiquettes lexicales, contrairement aux étiquettes grammaticales, peuvent prendre la forme d'un lemme quelconque du lexique cible), qui sont deux facteurs de complexité. En revanche, la réalisation d'une annotation au niveau des morphèmes contribue à limiter la diversité des unités en source, ce qui pourrait, au contraire, être un facteur facilitant.

2 Méthodologie

Modèles Pour développer un premier système, nous avons choisi de décomposer la tâche de génération de gloses en deux étapes : (1) génération d'une série d'étiquettes de gloses grammaticales (Premières Étiquettes, PE), n'utilisant que la phrase source (S) ; (2) génération des gloses restantes en s'appuyant sur la traduction (T).

L'étape (1) est effectuée avec un champ aléatoire conditionnel (*Conditional Random Field*, CRF) (Lafferty *et al.*, 2001), en utilisant Wapiti (Lavergne *et al.*, 2010). Afin de disposer d'un ensemble fini d'étiquettes à cette étape, toutes les gloses lexicales sont unifiées sous la même étiquette, « stem », suivant la méthodologie de (Moeller & Hulden, 2018; Barriga Martínez *et al.*, 2021).

L'étape (2) consistera alors à prédire les gloses lexicales pour spécialiser l'étiquette « stem » à partir de la traduction, par exemple en utilisant des alignements automatiques. Elle n'est pas traitée dans cette étude.

Ressources linguistiques Notre première langue d'étude est le tsez, une langue nakho-daghestanienne, en utilisant les annotations extraites de (Abdulaev & Abdulaev, 2010). Ce corpus a déjà été utilisé dans (Zhao *et al.*, 2020), qui aborde la même tâche avec une approche entièrement neuronale. Il comporte 2 000 phrases glosées et traduites en anglais, comme dans l'exemple 1.

Dans un second temps, nous avons également étudié le zaar [1], une langue tchadique parlée au Nigéria, à travers un corpus doublement annoté pour chaque morphème en glose et partie du discours (Caron, 2015). Nous dénombrons dans cette ressource 190 étiquettes grammaticales (sans stem) et 29 PoS.

1. Disponible sur https://github.com/surfacesyntacticud/SUD_Zaar-Autogramm/tree/master/CORPUS_PREANNOTE.

La figure 2 présente un exemple de phrase zaar avec ses deux types d'annotations au niveau des morphèmes, la strate P représentant ici les parties du discours.

S	Phrase segmentée	tò	kəndá	tɔ́	ndâː	lɔ́ːr–kɔ́nì	səmbɔ́r–sə	ɗi
G	Glose	well	then	3PL.AOR	start	bring–NMLZ	stranger–PL	CTP
P	Partie du discours	PART	ADV	AUX	VERB	VERB–SCONJ	NOUN–DET	PART
T	Traduction (EN)	*Well then they started bringing guests.*						

FIGURE 2 – Extrait du corpus en langue zaar (Caron, 2015)

Le tableau 1 présente le nombre de phrases (N_{sent}), le nombre d'occurrences (N_{token}) et de types (N_{type}) de *morphèmes*, le nombre de gloses grammaticales (N_{gram} ; sans « stem ») et d'étiquettes PoS (N_{PoS}) pour ces deux corpus.

	N_{sent}	N_{token}	N_{type}	N_{gram}	N_{PoS}
Tsez	2000	40229	1603	158	/
Zaar	1707	16957	1690	190	29

TABLE 1 – Statistiques générales des corpus tsez et zaar

3 Résultats préliminaires

Nous présentons une expérience pour chaque langue. Pour le tsez, nous évaluons la tâche (1) directement, tandis que les étiquettes en partie du discours permettent de comparer les tâches d'étiquetage en gloses et en PoS pour le zaar.

Dans les deux expériences, nous séparons le corpus en trois parties et faisons varier le nombre d'énoncés pour l'apprentissage, en maintenant 200 phrases pour le développement et 200 pour le test.

Enfin, comme nous avons accès à une segmentation en mots et en morphèmes pour les deux langues, le texte segmenté en entrée du CRF intègre les tirets dans la représentation graphique. Ceci permet indirectement d'exprimer une information de position sur les morphèmes en distinguant notamment les unités qui sont positionnées au début des mots de celles qui sont à l'intérieur des mots (elles débutent dans ce cas par le caractère « - »). Ce choix a été motivé par une expérience en amont pour laquelle cette information n'était pas présente et qui a conduit à des systématiquement moindres que ceux présentés ci-dessous.

Conditions expérimentales Nous avons utilisé le paramétrage par défaut de Wapiti [2]. L'algorithme d'optimisation utilisé (par défaut) est OWL-QN (*Orthant-Wise Limited-memory Quasi-Newton*, Andrew & Gao 2007), qui semble être la meilleure approche dans des situations avec peu de données (Lavergne *et al.*, 2010). Nous présentons principalement les résultats moyennés de deux lancers avec leur écart relatif, en considérant deux jeux de données ré-échantillonnés.

2. Détaillé sur https://wapiti.limsi.fr/manual.html.

3.1 Génération de gloses en tsez

Notre première expérience se concentre sur l'étape (1), l'étiquetage des gloses par un CRF. Dans le tableau 2, nous avons suivi l'évolution du score de correction [3] (*accuracy*) en fonction de la taille des données d'entraînement. À titre indicatif, (Zhao *et al.*, 2020) obtient, avec environ 1 600 phrases d'entraînement en tsez [4], une *accuracy* de 84 % avec le modèle statistique de (McMillan-Major, 2020) et 87 % avec leur modèle neuronal, sur la tâche de génération de gloses *entière* (donc 1 + 2 selon notre décomposition).

Deux patrons ont été comparés pour le CRF : l'un ne prend en compte que les unigrammes d'étiquettes, l'autre intégrant aussi des bigrammes, tous deux sur une fenêtre de cinq mots. Nous reportons aussi les résultats obtenus avec deux autres systèmes : `stem`, qui prédit toujours l'étiquette la plus fréquente, « stem », et `maj`, qui utilise les données d'entraînement pour obtenir l'étiquette majoritaire d'un morphème et prédit toujours cette étiquette pour toutes les occurrences de la base de test (« stem » est l'étiquette par défaut qui est utilisée pour tous les mots inconnus).

Taille entraînement	200	500	800	1000	1300	1600
`stem`	47,0 (0,6)	47,0 (0,6)	47,0 (0,6)	47,0 (0,6)	47,0 (0,6)	47,0 (0,6)
`maj`	83,6 (0,0)	84,0 (0,2)	84,0 (0,5)	84,1 (0,4)	84,1 (0,4)	84,2 (0,4)
Unigramme	**84,5** (1,7)	89,3 (1,1)	90,7 (0,0)	91,3 (0,3)	91,5 (0,8)	92,5 (0,1)
Bigramme	84,3 (1,7)	**89,7** (1,2)	**90,8** (0,3)	**91,6** (0,3)	**92,1** (0,4)	**92,8** (-)

TABLE 2 – Évolution de l'*accuracy* en fonction du nombre de phrases dans les données d'entraînement dans le corpus tsez (moyenne de deux lancers (écart type)).

Le système `stem` témoigne de la prépondérance de l'étiquette « stem » dans les données, représentant environ la moitié des étiquettes de référence. Le système `maj` quant à lui stagne autour de 84, ce qui le rend moins intéressant avec des données d'entraînement plus important.

Hormis des résultats similaires pour les deux patrons, nous pouvons voir qu'avec seulement 200 phrases d'entraînement, le modèle parvient à atteindre une correction élevée, avec en particulier un F-score de plus de 91 % pour l'étiquette « stem » (qui est l'étiquette majoritaire).

Taille entraînement	200	500	800	1000	1300	1600
`maj`	97,3	97,1	97,1	97,1	97,1	97,1
Unigramme	97,4	97,4	97,3	97,3	97,3	97,3
Bigramme	97,3	97,4	97,3	97,3	97,3	97,4

TABLE 3 – Évolution du taux de d'étiquettes « stem » correctement prédites en fonction du nombre de phrases tsez d'entraînement (résultat d'un lancer). Le système `stem` obtient par définition 100.

Le tableau 3 se concentre sur la prédiction de l'étiquette « stem » parmi les étiquettes de référence. Tous les systèmes obtiennent autour de 97 %, quel que soit le nombre de phrases en entraînement.

3. Proportion d'étiquettes correctement prédites.
4. Le nombre total de phrases dans leur corpus tsez est de 1 782 phrases.

Distinguer les étiquettes lexicales de celles grammaticales semblerait donc relativement accessible, avec peu de données.

Pour vérifier cette intuition, dans une expérience complémentaire, nous avons unifié toutes les étiquettes grammaticales sous le label « gram » et toutes les autres sous le label « stem ». Pour cette expérience, toujours pour le tsez, nous obtenons plus de 95 % d'*accuracy*.

3.2 Difficulté de la tâche par rapport à l'étiquetage en PoS

Avec la même méthodologie que pour la première expérience, nous évaluons aussi l'étiquetage en glose et en partie du discours pour le zaar. Les résultats sont présentés dans le tableau 4.

Taille	207 PoS / Gloses	507 PoS / Gloses	807 PoS / Gloses	1007 PoS / Gloses	1307 PoS / Gloses
`stem`	- / 44,7 (1,8)	- / 44,7 (1,8)	- / 44,7 (1,8)	- / 44,7 (1,8)	- / 44,7 (1,8)
`maj`	**71,6** (12,7) / **83,7** (9,8)	**80,5** (5,7) / **87,9** (5,1)	**83,9** (4,0) / **88,8** (4,5)	**85,1** (3,6) / **89,0** (4,2)	86,2 (3,7) / **91,1** (1,2)
Unigramme	61,8 (8,1) / 67,8 (0,3)	77,7 (4,2) / 79,3 (1,9)	82,1 (2,4) / 82,1 (2,3)	83,5 (3,7) / 83,8 (3,2)	85,7 (2,2) / 86,4 (1,2)
Bigramme	61,8 (8,6) / 67,9 (0,1)	77,6 (4,7) / 79,0 (2,6)	82,2 (2,4) / 82,3 (2,3)	83,1 (4,0) / 83,6 (3,4)	**86,2** (2,3) / 85,3 (-)

TABLE 4 – Évolution de l'*accuracy* en fonction du nombre de phrases dans les données d'entraînement dans le corpus zaar (moyenne de deux lancers (écart type)).

En ce qui concerne les deux patrons de CRF, nous observons des tendances similaires pour la génération de gloses par rapport à l'expérience précédente. Nous notons toutefois des *accuracy* plus faibles par rapport au tsez, sans doute explicables par le nombre réduit (d'occurrences) de morphèmes dans le corpus zaar. Cette différence peut aussi expliquer les performances du système `maj`, bien meilleur lorsqu'il y a peu de données (environ 15 points de différence avec 207 phrases d'entraînement).

De plus, nous pouvons voir que l'étiquetage en gloses obtient de meilleurs scores de correction par rapport à celui en PoS, bien que très proches. La différence est plus notable lorsque la taille de données est moindre. Malgré les facteurs de complexité énoncés, la tâche de génération de gloses semble donc de difficulté comparable à l'étiquetage en parties du discours.

4 Conclusions et perspectives

Les premiers résultats de génération de gloses sur des langues peu dotées semblent encourageants, avec une *accuracy* qui dépasse 80 % pour seulement 400 phrases à l'entraînement en tsez. Diverses perspectives sont envisagées : (a) étudier la difficulté de l'étape (2), qui semble comparativement plus simple ; (b) étendre le modèle par des ressources externes (dictionnaires, phrases non étiquetées) ; (c) construire des modèles de génération intégrés capables de réaliser simultanément les deux étapes.

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). Les auteurs remercient Bernard Caron pour la mise à disposition du corpus zaar et Antonios Anastasopoulos pour les données tsez.

Références

ABDULAEV A. K. & ABDULAEV I. K. (2010). *Cezjas fol'klor : (gíurus mecrek°iorno butirno) = Dido (Tsez) folklore = Didojskij (cezskij) fol'klor*. Leipzig : Lotos.

ANDREW G. & GAO J. (2007). Scalable training of l1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, p. 33–40, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/1273496.1273501.

BARRIGA MARTÍNEZ D., MIJANGOS V. & GUTIERREZ-VASQUES X. (2021). Automatic interlinear glossing for Otomi language. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, p. 34–43, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.americasnlp-1.5.

CARON B. (2015). Mettouchi, Amina, Martine Vanhove & Dominique Caubet (eds) (2012). 'The CorpAfroAs Corpus'. ANR CorpAfroAs : a Corpus for Afro-Asiatic languages. Document électronique. Esquisse grammaticale du zaar (langue tchadique du Nigéria), HAL : halshs-00647526.

LAFFERTY J. D., MCCALLUM A. & PEREIRA F. C. N. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, p. 282–289, San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 504–513 : Association for Computational Linguistics.

MCMILLAN-MAJOR A. (2020). Automating gloss generation in interlinear glossed text. In *Proceedings of the Society for Computation in Linguistics 2020*, p. 355–366, New York, New York : Association for Computational Linguistics.

MOELLER S. & HULDEN M. (2018). Automatic glossing in a low-resource setting for language documentation. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, p. 84–93, Santa Fe, New Mexico, USA : Association for Computational Linguistics.

ZHAO X., OZAKI S., ANASTASOPOULOS A., NEUBIG G. & LEVIN L. (2020). Automatic interlinear glossing for under-resourced languages leveraging translations. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 5397–5408, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.471.