Journées scientifiques "Linguistique informatique, formelle et de terrain" / Scientific meeting of the "Computational, formal and field linguistics" research group

28-29 nov. 2019
Orléans
France

Table des matières

Visualisation en arbres pour assister l'étude de la phonologie : application à la transcription de textes oraux nisvais, langue orale parlée dans l'ile de Malekula, au Vanuatu

Aznar Jocelyn [1]
(1) CREDO, adresse, 13001 Marseille, France
contact@jocelynaznar.eu

RÉSUMÉ

La visualisation de la transcription des mots en arbre a aidé l'analyse de la distribution des sons et des phonèmes du nisvai. Le retour critique que fournit la visualisation permet d'évaluer rapidement la cohérence de la transcription et facilite la conception d'hypothèses sur les processus phonologiques opérant dans la langue. Dans le cadre de l'étude du nisvai, cette visualisation a ainsi facilité l'identification des phonèmes, des structures syllabiques les plus courantes ou au contraire des hapax correspondant à des erreurs d'annotation ou des comportements particuliers.

ABSTRACT

Tree visualization to assist in the study of phonology : application to the transcription of Nisvai oral texts, an oral language spoken in Malekula Island, Vanuatu

Visualization of the transcription of words into trees helped in the analysis of the distribution of sounds and phonemes in nisvai. The critical feedback provided by visualization makes it possible to quickly assess the consistency of the transcription and facilitates the design of hypotheses on phonological processes operating in the language. In the context of the study of nisvai, this visualization thus facilitated the identification of phonemes, the most common syllabic structures or, on the contrary, hapaxes corresponding to annotation errors or particular behaviours.

MOTS-CLÉS : visualisation, corpus, arbre, phonologie, transcription.

KEYWORDS: visualisation, corpus, tree, phonology, transcription.

1 Une visualisation dans le cadre de l'étude du nisvai, une langue orale non documentée du Vanuatu

La visualisation en arbre de la transcription a été appliquée à l'étude de la phonologie du nisvai. Le nisvai est une langue orale parlée dans le sud-est de Malekula, au Vanuatu, par environ 200 locuteurs natifs. La langue avait fait partie d'une étude lexicographique (Charpentier, 1984) de 19 langues de la région. L'étude des pratiques narratives nisvaies a été réalisée en dialogue avec la demande de locuteurs du nisvai afin qu'une écriture du nisvai soit élaborée.

La transcription du nisvai a deux enjeux principaux : elle facilite l'étude des pratiques narratives nisvaies et elle sert de support à l'écriture du nisvai. Lors du terrain, la transcription a pu être effectuée avec les locuteurs et est alors devenu un entretien durant lequel le chercheur et les locuteurs ont pu

discuté de la pratique de la langue (Telban, 1997). Enfin, les différentes étapes de la transcription, qu'elle soit phonétique ou phonologique, ont correspondu à des étapes où le chercheur a pu poser des hypothèses sur les structures et les variations qu'il a observé.

Du point de vue du traitement informatisé des données, la transcription et les autres annotations des pratiques narratives nisvaies a été réalisée avec le logicel ELAN (Brugman & Russel, 2004). Les données linguistiques issues des annotations ont ensuite été intégrées dans une base de données relationnelle. Enfin, pour réaliser les visualisations en arbres, les données ont tout d'abord été extraites de la base puis mises en forme selon le langage DOT afin de produire des graphes, à l'aide du module python Graphviz (Ellson et al., 2004).

2 Les arbres des occurrences des caractères : un retour visuel sur ses analyses pour le linguiste

La visualisation des données en arbre proposée ici a été inspiré de la visualisation *Word Tree* (Wattenberg & Viegas, 2008). Il s'agit d'une représentation graphique interactive d'un texte où chaque mot peut être le point de départ d'un arbre et où chaque embranchement distingue les différentes occurrences à partir du mot sélectionnée.

La représentation en arbres des transcriptions du corpus permet d'explorer les données sous un angle différent de celui qu'offre un texte. Au lieu de voir les mots dans leur ordre d'apparition, le chercheur peut consulter les mots de son corpus en fonction de l'ordre d'apparition des caractères au sein d'un mot. Cette représentation permet d'observer les différents mots-formes possibles au sein d'un corpus. Les visualisations en arbres affichent également le nombre d'occurrences des formes étudiées au sein du corpus à chaque embranchement.

Ces représentations en arbres permettent au chercheur de critiquer ses propres analyses. Les données lui sont présentées sous un format différent de celui de l'annotation, facilitant un changement de perspective qui aide à explorer les structures existantes.

Le script qui a été développé pour proposer ces visualisations en arbres propose plusieurs options afin de paramétrer les arbres qui seront représenter. Une première possibilité est de produire une « forêt » d'arbres, résultant de l'analyse du corpus complet (voir la figure 1). Celle-ci est cependant trop vaste pour être intégré au sein d'un support papier, de plus le linguiste concentre souvent son analyse sur un ou quelques phonèmes uniquement. Le script offre aussi la possibilité de restreindre l'analyse aux mots-formes contenant un ou une suite de caractères (2). Enfin, il est également possible de catégoriser les caractères en classe. La figure 3 [1] fournit un exemple dans lequel les caractères correspondant à la classe des consonnes ont été catégorisé «C», les caractères correspondant à la classe des voyelles ont été catégorisés «V» et les caractères correspondant aux semi-consonnes ont été catégories «S».

Contrairement à la visualisation *Word Tree*, notre proposition ne permet pas d'interactions lors de la visualisation. Toutefois la production des arbres peut être contrôlée en sélectionnant au préalable les fichiers contenant les annotations qui seront intégrés dans la base de données linguistiques.

Cette communication sera l'occasion de voir comment les différentes visualisations en arbres des

1. Aucune méthode de segmentation automatique de la syllabe n'a été implémentée, la segmentation en syllabe du corpus a été fait manuellement pour produire ce graphe.

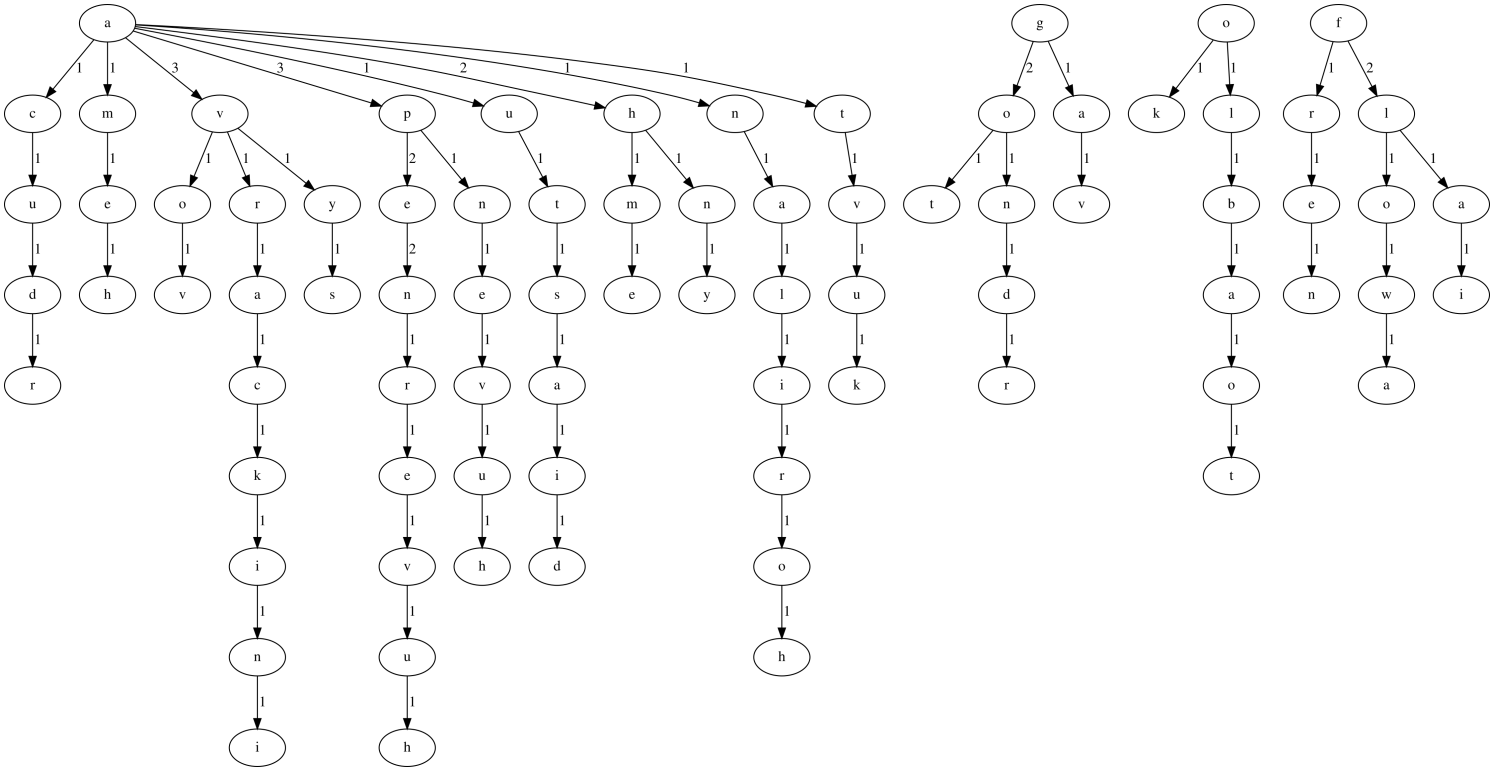


FIGURE 1 – Extrait des arbres de mots du corpus. Les nombres à côté des caractères indiquent le nombre d'occurrences de la lettre dans cette position.

occurrences des caractères au sein du corpus nisvai ont aidé à trouver des incohérences dans la transcription, identifier des processus phonologiques et des emprunt dans le corpus.

Références

Brugman, H. & Russel, A. (2004). Annotating multimedia/multi-modal resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (pp. 2065–2068). : Citeseer.

Charpentier, J.-M. (1984). *Atlas linguistique du Sud-Malakula (Vanuatu).* Coll. Langues et cultures du Pacifique, 2. SELAF.

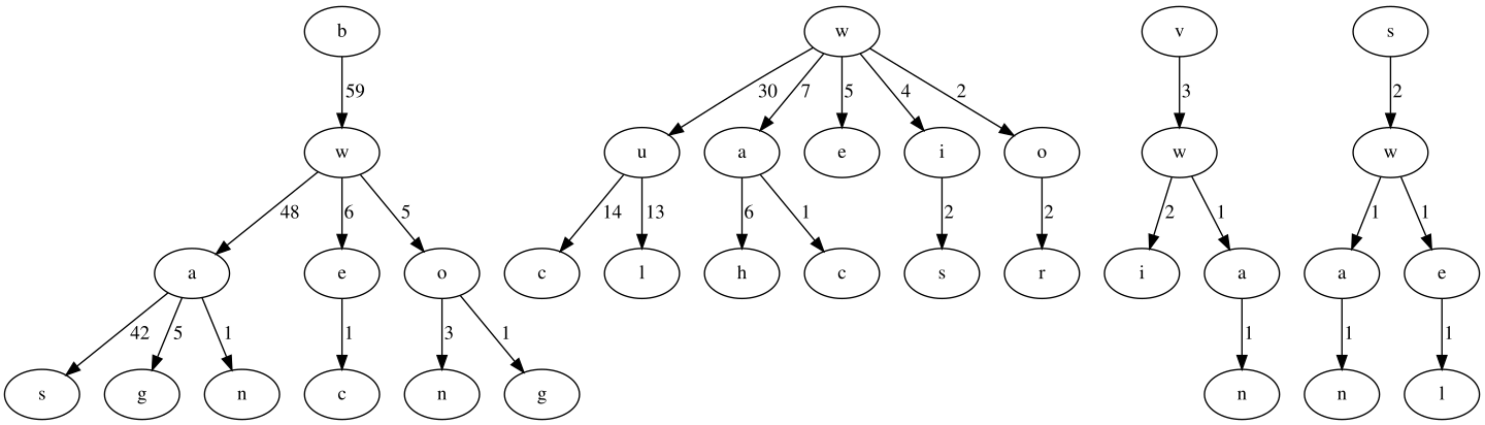


FIGURE 2 – Arbres pondérés des occurrences de la lettre «w» du corpus d'annotations des pratiques narratives nisvaies.

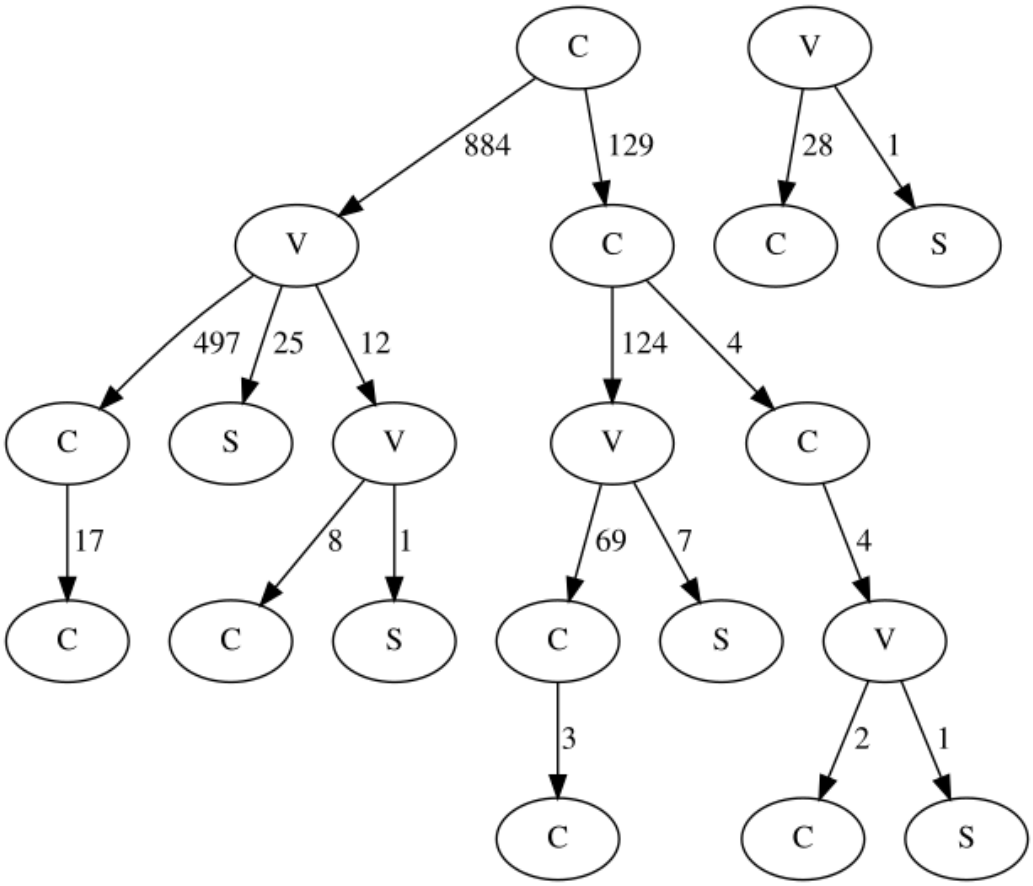


FIGURE 3 – Arbres des occurrences de syllabes nisvaies au sein du corpus.

Ellson, J., Gansner, E. R., Koutsofios, E., North, S. C., & Woodhull, G. (2004). Graphviz and dynagraph — static and dynamic graph drawing tools. In M. Jünger & P. Mutzel (Eds.), *Graph Drawing Software* (pp. 127–148). Springer Berlin Heidelberg.

Telban, B. (1997). Mutual understanding : Participant observation and the transmission of information in ambonwari. 20(1), 21–39.

Wattenberg, M. & Viegas, F. (2008). The word tree, an interactive visual concordance. 14(6), 1221–1228.

Analyse automatique du chinois utilisant des ressources linguistiques

Zhen CAI[1]

(1) Laboratoire ELLIADD, Université de Franche-Comté,30 rue Mégevand, 25000 Besançon

zhencai1122@hotmail.com

RESUME

Au cours de ces dernières années, les recherches sur l'analyse automatique de la langue chinoise se multiplient et différentes approches sont proposées afin d'améliorer ces analyses. Nous proposons dans cet article de construire un analyseur automatique utilisant des ressources linguistiques en taille réelle. Les problèmes à résoudre concernent le codage et le lexique : en particulier, nous avons construit un dictionnaire électronique avec la plateforme linguistique NooJ. Nous citons quelques exemples d'applications, parmi lesquelles un outil de segmentation automatique du chinois, essentiel pour construire toutes les applications de TAL pour le chinois.

ABSTRACT

In recent years, research on automatic Chinese language processing has exploded and different approaches are proposed to improve parsers. In this article, we present an automatic parser that uses large-coverage linguistic resources. The two main problems to solve are managing the encoding system and the Chinese lexicon. We have constructed a large-coverage electronic dictionary, using the NooJ platform. We present various applications, among them an automatic Chinese tokenizer, crucial for many NLP applications.

MOTS-CLES : chinois, codage, segmentation, dictionnaire
KEYWORDS : Chinese, encoding, Tokenization, Dictionary

---

1 Le traitement automatique du chinois

Depuis les années 90, les chercheurs en TAL ont commencé à construire des logiciels capables d'analyser automatiquement des textes chinois. Plusieurs problèmes doivent être abordés :

— Les problèmes de codage des textes

— La segmentation du chinois, étape cruciale pour toute application de TAL.

— Le traitement des morphèmes et mots composés

— Le traitement de certains objets linguistiques spécifiques aux langues asiatiques, tels que les classificateurs et les subordinateurs.

2 Problème de codage des textes chinois

Pour que les caractères chinois puissent être traités et affichés correctement, il faut normaliser le codage de caractères. Nous utilisons la plateforme NooJ qui traite les textes en Unicode (UTF8).

3 Unité linguistique du chinois

La première étape de toute analyse d'un texte chinois est de reconnaître ses unités lexicales. Pour cela, nous avons construit un dictionnaire électronique qui contient environ 63 000 entrées. Nous présentons les critères que nous avons utilisés pour définir ce que sont ces unités lexicales, un problème crucial en particulier lorsque nous avons affaire à des mots simples et à des mots composés (quand doit-on traiter en bloc une séquence de caractères). Noter que le problème de la segmentation est différent de celui pour les autres langues ; par exemple, en français, les mots composés se différencient des mots simples grâce à la présence de séparateurs comme l'espace ou le trait d'union, ce qui n'est pas le cas pour le chinois.

4 La segmentation du chinois

Nous proposons de comparer les outils de la segmentation du chinois existant (utilisant des techniques probabilistes ou statistiques) avec notre outil de segmentation du chinois (utilisant des dictionnaires et des grammaires). Nous commenterons les résultats et erreurs produits par notre outil.

5 Conclusion

Après une présentation succinte de l'état de l'art dans le domaine de la segmentation automatique de textes chinois, nous présenterons un outil de segmentation utilisant des ressources linguistiques sous la forme d'un dictionnaire électronique de 63 000 entrées et de grammaires locales, morphologiques et syntaxiques. Nous commenterons enfin les résultats produits par notre outil.

Références

Feng, Z.冯志伟 (2004). 机器翻译研究 Jīqì Fānyì Yánjiū 'Etudes sur la traduction automatique' China Translation & Publishing Corporation.

Guo,R. 郭锐(2002). 现代汉语词类研究 Xiàndài Hànyǔ Cílèi Yánjiū 'Etudes sur les catégories en chinois moderne'. Commercial Press商务印书馆.

Li, C. and Thompson, S. (1989). MANDARIN CHINESE A Functional Reference Grammar. UNIVERSITY OF CALIFORNIA PRESS.

Paris, M. (1981). Problème de syntaxe et de sémantique en linguistique chinoise. INSTITUT DES HAUTES ETUDES CHINOISES.

Paris, M. (1996). La subordination en chinois standard : quelques contraintes d'agencement. Livre « Dépendance et intégration syntaxique » Edité par Claude Muller P233-P240

Paris, M. (2007). Un aperçu de la réduplication nominale et verbale en mandarin. Livre « La réduplication » par Alexis Michaud et Aliyah Morgenstern. Edition Ophrys. P63-P76.

Paris, M. (2013). Linguistique chinoise et linguistique générale. L'Harmattan.

Silberztein, M. (2015). La formalisation des langues l'approche de NooJ. Editions ISTE.

Yang-Drocourt, Z. (2007). Parlons chinois. L'Harmattan.

A bioinformatics solution to inter-rater agreement for forced time-alignment of data from underresourced languages

Matthew Stave, [1] François Delafontaine, [1] Frank Seifart, [2] Ludger Paschen, [2]
(1) DDL, 14 avenue Berthelot, 69363 Lyon, France
(2) ZAS, Schützenstrasse, 10117 Berlin, Deutschland
matthew.stave@cnrs.fr, francois_delafontaine@outlook.com,
frank.seifart@berlin.de, paschen@leibniz-zas.de

RÉSUMÉ

Une solution de la bioinformatique à l'accord inter-annotateurs.
Un alignement précis des séquences d'annotations est un prérequis à l'analyse phonologique, et un aspect important de la linguistique de l'oral. Le projet DoReCo exploite l'alignement phonémique produit par le logiciel MAUS (Kisler et al. 2017) et, pour en évaluer la précision, a mesuré l'accord inter-annotateurs entre cet alignement automatique et des alignements manuels, ce qui pouvait impliquer l'ajout, le retrait ou l'altération d'unités d'annotation. Cette situation se révèle très problématique pour l'accord inter-annotateurs. L'algorithme de Needleman-Wunsch, du champ de la bioinformatique, offre une solution pratique et puissante. Son implémentation, comparée à une correction manuelle, connecte plus de 95% des unités correctement. L'algorithme offre une précision nouvelle pour la mesure de l'accord inter-annotateurs, et s'applique à d'autres tâches où cette mise en relation est requise.

ABSTRACT

Precise time-alignment for sequences of annotations is a prerequisite for phonological analysis, and an important aspect of oral linguistics. The DoReCo project relies on phonemic time-alignment by MAUS software (Kisler et al. 2017) and, to evaluate its precision, must measure inter-rater agreement between MAUS-aligned and manually-aligned segments, which involve adding, removing, and changing annotation units. This situation proves highly problematic for inter-rater agreement. The Needleman-Wunsch algorithm, from the bioinformatics field, offers a practical and powerful solution to that problem. Its implementation, when compared with a manual correction, matched over 95% of all units correctly. The algorithm offers a newfound precision for inter-rater agreement measurement, and has further applications where precise matching is required.

MOTS-CLÉS : Linguistique de corpus ; linguistique de l'oral ; accord inter-annotateurs ; Needleman-Wunsch ; bioinformatique.

KEYWORDS: Corpus linguistics ; oral linguistics ; inter-rater agreement ; Needleman-Wusch ; bioinformatics.

1 Introduction

We propose an innovative solution to a well-known problem in the area of inter-rater agreement, regarding the problem of measuring segmentation agreement on annotations which differ in content

and/or number of units (Mathet et al. 2015). This solution from bioinformatics, called the Needleman-Wunsch algorithm (Needleman & Wunsch 1970), allows for the pairing of annotation units and, from there, the precise measurement of agreement.

First we will present the context of this work, that is, the DoReCo project and its specific inter-rater agreement task (point 2). We will then review some methods to address that task (point 3) before presenting the Needleman-Wunsch implementation and its results (point 4).

2 An alignment problem in the DoReCo project

DoReCo[1] is a new French-German collaborative project designed to bring together spoken language corpora from 50+ languages, taken from documentations of small and often endangered languages. This is done both to bring awareness to lesser-studied languages and language communities, and to enable access to a more diverse sample of languages for researchers to test linguistic hypotheses. The DoReCo project itself will use the corpora to explore a number of questions related to universal claims made about language production in articulatory phonetics and information-processing.

One important contribution of the project is the time-alignment of all the transcriptions, using the MAUS time-alignment software (Kisler et al. 2017). This is necessary for answering questions of phonetic (in)compressibility and final lengthening. Using MAUS, transcriptions are time-aligned at the word and phoneme level, using a global phonemic-alignment model, and these aligned corpora are then made available to the public.

Due to this reliance on the MAUS alignment, it was necessary to test the accuracy of the obtained phonemic tiers. This was to be done by comparing MAUS's segment boundaries with segment boundaries from manually-corrected tiers, to determine the agreement between word and phoneme boundaries. However, measuring agreement between sequences with different time boundaries (segmentation), as opposed to different content (categorization), is known to be problematic (Mathet & Widlöcher 2016). As it turns out, the manual corrections of the MAUS alignments can involve adding, removing, or altering the content of segments, thus rendering the comparison uncertain in multiple ways. Without knowing which two segments to compare, assessing the accuracy of the MAUS alignment is not possible.

3 Existing tools & methods

The standard method for evaluating categorical agreement of items between annotators is Cohen's kappa (Cohen 1960, 1968), which calculates the proportion of observed agreement and normalizes it by the predicted chance agreement. Other methods, such as boundary distance, can be used to evaluate continuous agreement of temporal boundaries, but these methods require that items be properly aligned with corresponding items on the other tier(s). When tiers are not aligned (e.g. after insertions or deletions), it becomes necessary to first determine this alignment between the tiers to be compared. Without this alignment, at best the mean distribution can be obtained, as for example with Krippendorff's alpha (Krippendorff 1970).

This task of aligning segments is often done manually, but when working with hundreds of thousands of annotation units across thousands of files, as in the DoReCo corpora, this kind of manual alignment becomes unfeasible. There has been some work on the development of automated

[1] <http://doreco.info>

methods of segment alignment. Holle & Rein (2015) have created the EasyDIAG tool, which uses categorical (same tier type and same label) and continuous (percent overlap) approaches to align segments. The STACCATO algorithm (Lücking et al. 2011) uses mutual overlap of multiple annotators to determine "nuclei" segments. Other approaches ignore alignment altogether, simply reporting the raw amount of overlap of any segment with another segment, as a proportion of the segment lengths, averaged across all segments (Strunk et al. 2014).

A method called Gamma (Mathet et al. 2015) has been specifically devised for this task. It uses a unified approach whereas pairing units and measuring their agreement is done parallel to each other, in one process. The result is not only a measurement, but also an automatically realigned annotation. It offers a good alternative to the most common method to our knowledge, which is atomization (*idem* : 440 ; Krippendorff 2004), or the reduction of the segmentation problem to a categorization one by segmenting the annotations further down into intervals of equivalent size. Atomization then allows for a simple Kappa-score, making for an easily sharable measurement.

These methods, however, while often quite useful for comparing aligned segments, proved inadequate for aligning MAUS sequences with manually-corrected sequences. This led us to explore approaches to sequence alignment in other fields, such as bioinformatics.

4 A bioinformatic solution

The Needleman-Wunsch algorithm (Needleman & Wunsch 1970) is a well-known algorithm in bioinformatics, where it is frequently used in aligning amino acid sequences in proteins, or nucleotide sequences in DNA strings. It has the advantage of returning the optimal alignment(s) of two sequences of items, based only on their labels and their positions in the sequences. With very little adjustment (here our labels are word or phoneme strings, rather than amino acids), this can be employed on linguistic annotations to find the optimal alignment of two sequences of words, phonemes, or other annotations.

The algorithm works by maximizing the similarity between the two sequences, allowing for three edit operations: insertions, deletions, and substitutions. Each of these edit operations can be weighted differently, depending on one's theoretical considerations (for our study we have weighted all three equally). A matrix (length(seqA) by length(seqB)) is created, populated by the similarity score for every pairwise combination in the two sequences: identical items receive +2, while insertions, deletions, and substitutions receive -1. This generates a set of optimal paths from the first pair of items to the final pair of items. The alignment paths with the highest score are returned. Within the DoReCo project, the implementation of these algorithm is done using a Python library called *biopython*, and its "pairwise2" function. A preliminary test was done on transcriptions extracted from ELAN tiers in three languages: Anal (India), Resigaro (Colombia), and Vera'a (Vanuatu), with a total of 2503 words.

To test the method, a Cohen's kappa was performed on the Needleman-Wunsch-aligned sequence and a manually-aligned sequence. The reliability of the algorithm compared to human matching of sequences resulted in a kappa-score of 0.97-1, with 95-99% of sequences being aligned the same way across the files: most of the divergence is due to misinterpreted pauses. At that level of reliability, and for the purpose of inter-rater agreement, this suggests the method can be entirely automated. As for the results it gave: we were notably able to establish mean differences in unit onsets and offsets, meaning how much each unit boundary, start and end, was moved, in

(milli)seconds, as well as the proportion of moved boundaries and as such, the amount of work a manual correction required. We were also able to reliably track substitutions, additions and deletions due to corrections.

5 Perspectives

Preliminary results have been very promising. Further work will include expanding the sample of test languages, and of course making use of the aligned segments to assess the accuracy of the MAUS word and phoneme alignment. Which method to employ for this latter task is still under consideration.

We do however see applications even beyond inter-rater agreement. This pairing will, as examples, help with operations such as the merging of different transcription files, or the realignment of morphemic units under their word-corrected counterpart. While such tasks do require some manual correction of the automatic pairing to be fully effective, they are made possible by this very bioinformatics method.

References

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20(1), 37-46.

Cohen, J. (1968). Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* 70(4), 213-220.

Holle, H. & Rein, R. (2015). EasyDIAg: A tool for easy determination of interrater agreement. *Behavior research methods* 47(3), 837-847.

Lücking, A., Ptock, S. & Bergmann, K. (2011). Assessing agreement on segmentations by means of Staccato, the SegmenTation Agreement CalCulator According to Thomann. *Proceedings of the 9th international conference on Gesture and Sign Language in Human-Computer Interaction and Embodied Communication*. <researchgate.net/publication/262171036_Assessing_Agreement_on_Segmentations_by_Means_of_Staccato_the_Segmentation_Agreement_Calculator_according_to_Thomann>.

Kisler, T., Uwe, R.D. & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer speech & language* 45, 326-347.

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and psychological measurement* 30, 61-70.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Thousand Oaks: Sage.

Mathet, Y., Widlöcher, A. & Métivier, J.-P. (2015). The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational linguistics* 41(3), 437-479.

Mathet, Y. & Widlöcher, A. (2016). Évaluation des annotations : ses principes et ses pièges. *TAL* 57(2), 73-98.

Needleman, S.B. & Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology* 48(3), 443-453.

Strunk, J., Schiel, F. & Seifart, F. (2014). Untrained forced alignment of transcriptions and audio for language documentation corpora using WebMAUS. *LREC 2014*, Reykjavik, Iceland.

ACCOLÉ : Annotation Collaborative d'erreurs de traduction pour COrpus aLignÉs

Emmanuelle Esperança-Rodier[1] Francis Brunet-Manquat[1]
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP*, LIG, 38000 Grenoble, France

emmanuelle.esperanca-rodier@univ-grenoble-alpes.fr, francis.brunet-manquat@univ-grenoble-alpes.fr

RÉSUMÉ

Cet article présente une plateforme pour l'édition collaborative d'erreurs de traduction. Cette plateforme, nommée ACCOLÉ (Annotation Collaborative d'erreurs de traduction pour COrpus aLignÉs), propose une palette de services permettant de répondre aux besoins d'analyse d'erreurs de traduction : gestion simplifiée des corpus et des typologies d'erreurs, annotation d'erreurs, collaboration et/ou supervision lors de l'annotation, recherche de modèle d'erreurs dans les annotations. Dans cet article, nous présentons un état de l'art des services d'annotation, puis nous décrivons la plateforme ACCOLÉ et les corpus annotés obtenus jusqu'ici.

ABSTRACT

ACCOLÉ: A Collaborative Platform of Error Annotation for Aligned Corpora. This article presents a platform for the collaborative editing of translation errors. This platform, named ACCOLÉ, offers a range of services that meet the analysis needs of translation errors: simplified management of corpora and typologies of errors, annotation of errors, collaboration and/or supervision during annotation, looking for error types in annotations. In this article, we present a state of the art of the annotation services, we then describe the platform ACCOLÉ and the annotated corpora obtained so far.

MOTS-CLÉS : Annotations d'erreurs de Traductions Automatique, Annotation collaborative, Evaluation de la qualité de la TA

KEYWORDS: Annotations of translation errors, Collaborative annotation, Machine Translation Quality Assessment

1 Contexte

L'analyse d'erreurs de traduction est une tâche qui s'est développée ces dernières années afin, d'une part, d'améliorer les systèmes de Traduction Automatique (TA) et d'autre part d'évaluer qualitativement ces mêmes systèmes. Outre le domaine de la TA, l'analyse d'erreurs de traduction est notamment utilisée afin d'améliorer les méthodologies d'enseignement en traduction (Kübler et al, 2016). L'intégration de la TA dans les masters de traduction apporte un questionnement nouveau quant à l'évaluation des systèmes de TA. Quel est le ressenti des traducteurs face aux différents systèmes de TA (Esperança-Rodier et al., 2017) ? Quels outils utiliser pour les non spécialistes ? La réponse à ces questions nécessite une analyse linguistique des erreurs de TA afin de pouvoir,

premièrement, révéler les erreurs typiques de chaque type de systèmes de TA, deuxièmement, corréler ces erreurs typiques à la capacité de post-édition des utilisateurs, et troisièmement créer des corpus. Réalisée manuellement, l'analyse d'erreurs de traduction est une tâche très chronophage. Des recherches portent donc sur l'analyse automatique d'erreurs de traduction (Popović et al., 2006). Manuelle ou automatique, cette tâche requiert des critères d'analyse des erreurs, souvent liés aux caractéristiques particulières de l'outil utilisé. Les travaux de Vilar et al. (2006), du projet MeLLANGE (Castagnoli et al., 2011), de Felice et al. (2012), Wisniewski et al. (2013.a) ainsi que les travaux de Lommel (2018) portent sur la définition de typologies d'erreurs au niveau linguistique. Une typologie d'erreurs consiste en une liste d'erreurs, catégorisées en fonction des actions de corrections à effectuer pour améliorer la traduction. Par exemple, le type d'erreurs "mot manquant" indique qu'un mot du segment source n'a pas été traduit dans l'hypothèse de traduction et donc n'y figure pas. Afin d'améliorer l'hypothèse de traduction, il faut mener l'action corrective de traduire ce "mot manquant". ACCOLÉ propose ces différentes typologies d'annotation d'erreurs, mais permet également de proposer de nouvelles typologies, mieux adaptées à la corrélation des erreurs des systèmes aux erreurs non corrigeables par les utilisateurs. D'autres travaux portent sur la création de corpus d'erreurs de traduction et de correction d'erreurs de traduction comme Potet et al. (2012) et (Wisniewski et al., 2013.b). D'autres recherches se concentrent sur la création d'outils permettant soit l'estimation de la qualité comme QuEst++ (Specia et al., 2015), ou l'évaluation de la qualité avec Kantan's MT LQR et PET (Aziz et al. 2012) soit l'analyse d'erreurs tels que BLAST (Stymne, 2011), focalisé sur l'annotation, Coreference Annotator (Tsoumari et al., 2011) plus orienté sur les corpus parallèles, YAWAT (Germann, 2008) un outil d'alignement qui autorise l'étiquetage, et TranslationQ (Steurs, 2015) orienté traduction et révision. C'est dans cette dernière veine que nous nous positionnons. Nous voulons analyser les erreurs de traduction sur différents corpus, en collaboration avec des personnes de formations diverses. La tâche d'analyse des erreurs étant déjà fastidieuse, nous avons créé une plateforme accessible en ligne (http://lig-accole.imag.fr). Dans la suite de cet article, nous présenterons notre plateforme ACCOLÉ, avant de présenter les corpus annotés obtenus jusque là.

2 ACCOLÉ, une plateforme pour l'annotation d'erreurs

2.1 Présentation

L'objectif initial qui a guidé le développement d'ACCOLÉ, est l'annotation manuelle des erreurs de traduction selon des critères linguistiques. L'idée sous-jacente est de pouvoir fournir à un utilisateur une aide dans le choix d'un système de TA à utiliser selon le contexte (compétences linguistiques et informatiques de l'utilisateur, connaissance du domaine du document source à traduire et la tâche pour laquelle il a besoin de traduire le document source.) Pour ce faire, ACCOLÉ doit permettre de détecter quels sont les phénomènes linguistiques qui ne sont pas traités correctement par le système de TA étudié. Nous proposons sur la même plateforme une palette de services permettant de répondre aux besoins d'analyse d'erreurs de traduction. Ainsi, les principales fonctionnalités de la plateforme ACCOLÉ sont la gestion simplifiée des corpus, des typologies d'erreurs, des annotateurs, etc. ; l'annotation d'erreurs ; la collaboration et/ou supervision lors de l'annotation ; la recherche de modèles d'erreurs (type d'erreurs dans un premier temps, patrons morphosyntaxiques ultérieurement) dans les annotations. La tâche d'analyse d'erreurs étant déjà fastidieuse, il est important que les personnes la réalisant aient un accès simple à l'outil ainsi qu'au corpus qu'ils souhaitent analyser. La plateforme est donc disponible en ligne (http://lig-accole.imag.fr) sur un navigateur et ne nécessite aucune installation spécifique.

2.2 Gestion des projets d’annotations

Une tâche d’annotation est décrite sur la plateforme par la notion de projet, soit un couple constitué d’un corpus et d’une typologie d’erreurs. Ainsi, un corpus pourra être associé à plusieurs typologies d’erreurs sous forme de plusieurs projets d’annotations (le corpus ne sera alors chargé qu’une fois sur la plateforme). Le responsable d’un projet fournira également la liste des annotateurs et des superviseurs. La gestion des typologies d’erreurs est réalisée par les responsables de la plateforme. Un type d’erreur sera composé d’un nom, d’une catégorie (facultatif), d’une sous-catégorie (facultatif) et d’un code (raccourci clavier pouvant être utilisé lors de l’annotation). Des corpus, par exemple des extraits du BTEC (Basic Travel Expression Corpus) ou bien de nouvelles journalistiques, ainsi que la typologie d'erreur de Vilar et DQF-MQM sont déjà disponibles sur ACCOLÉ. De plus, la plate-forme permet de téléverser de nouveaux corpus et de saisir d'autres typologies d'erreurs.

2.3 Annotation d’erreurs

La plateforme ACCOLÉ propose de visualiser et d’annoter les erreurs d’un couple de phrases source/cible. La figure 1 présente l’interface proposée à l’annotateur. L’annotation se fait en deux

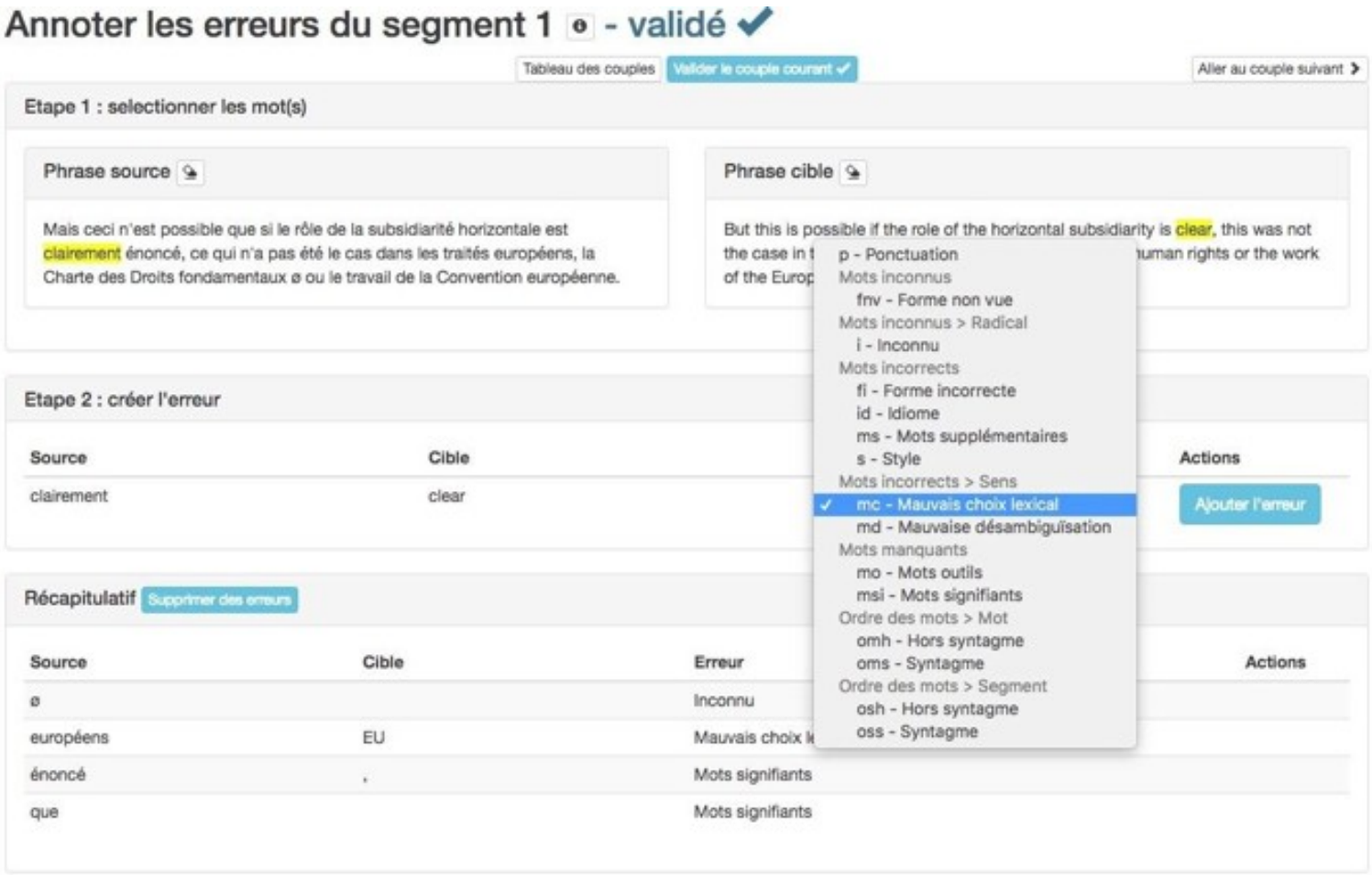


FIGURE 1 : annoter une erreur sur la plateforme ACCOLÉ avec la typologie créée par Vilar et al. (2006)

étapes. La première étape consiste à sélectionner, à l’aide de la souris, des mots dans la phrase source, et de leur équivalent dans la phrase cible, présentant une erreur de traduction. La seconde étape consiste à choisir le type d’erreurs soit à l’aide de la souris, soit à l’aide des raccourcis clavier,

à associer au couple des mots sources/cibles préalablement sélectionnés. Il est important d'associer le type d'erreurs à la fois à la source et à la cible afin de dégager des modèles d'erreurs qui permettront in fine de propager les annotations sur d'autres corpus.

Pour répondre aux problèmes d'accord inter-annotateurs (Popović, 2018), ACCOLÉ propose deux mécanismes pour aider l'annotateur dans sa tâche : un mécanisme de supervision permettant à un responsable de contrôler l'avancée de la tâche, ce mécanisme encourage surtout la communication entre superviseur et annotateur par la possibilité de créer des fils de discussion pour un couple de phrase source/cible précis (demander des précisions sur un type d'erreurs, pointer une erreur d'annotation, etc.); et un mécanisme collaboratif permettant aux annotateurs de communiquer autour d'un couple phrase source/cible précis (ce mécanisme est une option à activer dans le projet).

2.4 Représentation des erreurs basée sur les SSTC

La plateforme utilise une représentation des données base e sur les SSTC (Structured String-Tree Correspondences, Boitet et Zaharin 1988). Une erreur est constituée d'une étiquette et d'un ensemble de SNODE (intervalle représentant la sous-chaîne dans la phrase source ou cible correspondante). Par exemple dans la figure 2, l'erreur portant sur "toute l'" et "any" est décrite par son étiquette Mauvais choix lexical (cat. Mot incorrect, sous-cat. Sens), par son positionnement dans la phrase source (SNODE [49-56] - sous chaîne entre le 49ème caractères et le 56ème) et la phrase cible (SNODE [46-48]). L'avantage d'utiliser ainsi les SNODE est de se passer d'une structure syntaxique pour décrire l'erreur.

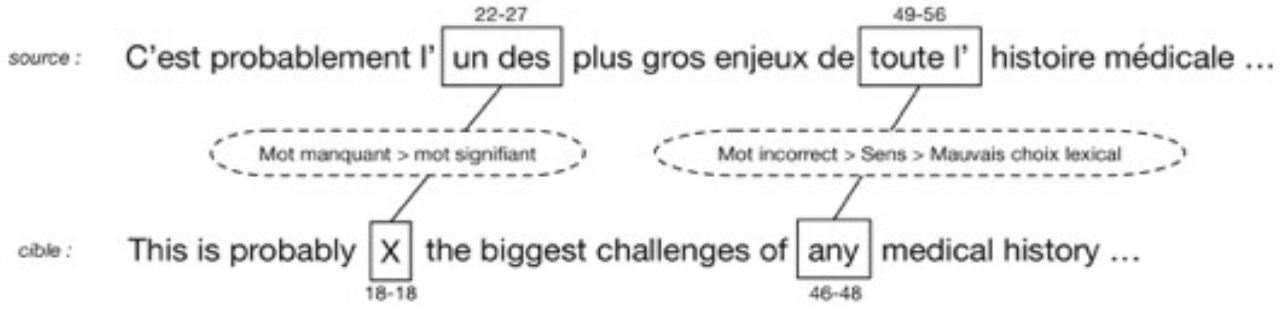


FIGURE 2 : exemple d'annotations

L'autre avantage est de pouvoir ajouter a posteriori des analyses morphosyntaxiques sur les phrases sources et cibles. Une ou plusieurs analyses (Stanford Tagger, Xerox Incremental Parser, etc.) pourront ainsi être rattachées à l'aide des SNODE aux phrases. A la fin de la tâche d'annotation, ses analyses serviront à rechercher des modèles d'erreurs (patrons morphosyntaxiques par exemple). L'idée est donc d'utiliser les erreurs comme une représentation pivot dans le mécanisme de recherche.

2.5 Corpus

À l'heure à laquelle nous écrivons cet article, ACCOLÉ propose 2 typologies d'erreurs (celles de Vilar et al. (2006) et DQF-MQM (Lommel, 2018)) ainsi que 15 corpus français-anglais (allant des

nouvelles journalistiques, à des documents techniques, des brevets, des textes monolingues français, des extraits du BTEC (Basic Travel Expression Corpus), jusqu'à des documents sur le climat) ayant permis la création de 19 projets. Ceux-ci correspondent à 6 817 phrases, 134 273 mots sources, 114 511 mots cibles, pour 23 525 annotations réalisées par des annotateurs natifs anglais. Ces corpus sont structurés selon les SNODEs et sont disponibles sur demande au format XML ou JSON. Une fonction permet de rechercher dans ces corpus les types d'erreurs en fonction des typologies utilisées. Au moment de la rédaction, nous travaillons sur la recherche de modèle d'erreurs.

3 Conclusion

Dans cet article, nous présentons la plateforme d'ACCOLÉ. Nous décrivons la représentation interne des données et ses avantages. Nous souhaitons aussi l'améliorer en permettant également l'association d'arbres de dépendances. Certains corpus annotés disponibles ont déjà été utilisés pour une comparaison linguistique de la qualité de la traduction de différents systèmes de TA (Esperança-Rodier et Becker, 2018). Une étude est prévue afin d'utiliser ACCOLÉ pour l'annotation bilingue d'EPL. La recherche de modèle d'erreurs dans les annotations, atout essentiel d'ACCOLÉ, nous permettra dans une prochaine version, de nous baser sur les modèles d'erreurs afin de propager les annotations d'erreurs de manière semi-automatique à de nouveaux corpus.

Références

Aziz, W., Sousa, S.C.M., Specia, L. (2012) PET: a tool for post-editing and assessing machine translation. In: *Calzolari N, Choukri K, Declerck T, Dogan MU, Maegaard B, Mariani J, Moreno A, Odijk J, Piperidis S (eds) Proceedings of the eighth international conference on language resources and evaluation*, Istanbul, pp 3982–3987.

Boitet, C. and Zaharin, Y. (1988). Representation trees and string- tree correspondences. In *Proceedings of international Conference on Computational Linguistics* COLING-88, 59-64.

Castagnoli, S., Ciobanu, D., Kübler, N., Kunz, K., Volanschi, A. (2011). Designing a Learner Translator Corpus for Training Purposes. In *Corpora, Language, Teaching and Resources: From Theory to Practice*. Edited by Kübler N. Bern: Peter Lang.

Esperança-Rodier, E. and Becker, N. (2018). Comparaison de systèmes de traduction automatique, probabiliste et neuronal, par analyse d'erreurs. Proceedings of the 4th day on « *Traitement Automatique des Langues et Intelligence Artificielle » - TALIA 2018 Day of the plate-forme Intelligence Artificielle (PFIA 2018)*. Nancy, France, 6 juillet 2018. Edited by Didier Schwab et Pierre Zweignebaum.

Esperança-Rodier, E., Rossi, C., Bérard, A., Besacier, L. (2017). Evaluation of NMT and SMT systems: A study on uses and perceptions. In *Proceedings of the 39th Conference Translating and the Computer,* Nov 2017, Londres, United Kingdom. Translating and the Computer 39.

Felice, M., Specia, L. (2012). Linguistic Features for Quality Estimation. In P*roceedings of the 7th Workshop on Statistical Machine Translation*, Montréal, Canada, June 7-8, 2012 Association for Computational Linguistics, 96–103

Germann, Ulrich. (2008) Yawat: Yet Another Word Alignment Tool. *46th Annual Meeting of the Association for Computational Linguistics.* June 15-20, 2008. Columbus, Ohio, USA. Demo Papers.

Kübler, N., Zimina, M., Fleury, S. (2016). Origines des erreurs en Traduction Spécialisée : différentiation textométrique grâce aux corpus de textes cibles annotés. *Actes de la conférence conjointe JEP-TALN-RECITAL 2016*, volume 09 : ELTAL, Paris.

Lommel, A., and Alan, K. M. (2018). Tutorial: MQM-DQF: A Good Marriage (Translation Quality for the 21st Century). *13th Conference of the Association for Machine Translation in the Americas (Volume 2: User Papers). Vol. 2.*

Popović, M., Hermann, N., de Gispert, A., Mariño, J.B., Gupta, D., Marcell, F., Lambert, P., and Banchs. R. (2006) Morpho-syntactic Information for Automatic Error Analysis of Statistical Machine Translation Output. *Workshop on Statistical Machine Translation.* 1-6. New York City. June 2006 ©2006 Association for Computational Linguistics

Popović, M. (2018) Error Classification and Analysis for Machine Translation Quality Assessment. *Moorkens J., Castilho S., Gaspari F., Doherty S. (eds) Translation Quality Assessment. Machine Translation: Technologies and Applications, vol 1. Springer, Cham*

Potet, M., Esperança-Rodier, E., Besacier, L., Blanchon, H. (2012). Collection of a large database of French-English SMT output corrections. In *CHAIR, N. C. C., CHOUKRI, K., DECLERCK, T., DO˘GAN, M. U., MAEGAARD, B., MARIANI, J., ODIJK, J. et PIPERIDIS, S., éditeurs : Actes de Eight International Conference on Language Resources and Evaluation* (LREC'12), Istanbul, Turkey. European Language Resources Association (ELRA).

Specia, L., Paetzold, G. H. et Scarton, C. (2015) Multi-level Translation Quality Prediction with QuEst++. *53rd Annual Meeting of the Association for computational linguistics and the 7th Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing.* System Demonstrations. Beijing, China. 115-120.

Steurs, F., Segers, W. and Kockaert. H. (2015) Translation Expert (TranslationQ & RevisionQ): Automated translation process with real-time feedback & evaluation/ revision with PIE. (Abstract No. Keynote speech 4). *Talking to the World, University of Newcastle,* UK. 09 Sep 2015-11 Sep 2015.

Stymne, S. (2011) Blast: A tool for error analysis of machine translation output. *The 49th Annual Meeting of the Association for Computational Linguistics.* System Demonstrations. 56–61. Portland, OR, USA, Jun.

Tsoumari, M. and Petasis, G. (2011) Coreference Annotator A new annotation tool for aligned bilingual corpora. *Recent Advances in Natural Language Processing.* 43–52. Hissar, Bulgaria. 12-14 September 2011.

Vilar, D., Xu, J., D'Haro, L.F. and al. (2006) Error analysis of statistical machine translation output. *5th International Conference on Language Resources and Evaluation.* 97-702.

Wisniewski, G., Singh, A.K., Sega,N. and Yvon, F. (2013.a) Un corpus d'erreurs de traduction. *TALN*

Wisniewski, G., Singh, A.K. and Yvon, F. (2013.b) Quality Estimation for Machine Translation: Some Lessons Learned, *MT Journal*

La Collection Pangloss

Séverine Guillaume[1] Alexis Michaud[1] Balthazar Do Nascimento[1]

(1) Langues et civilisations à tradition orale, 7 rue Guy Môquet, 94801 Villejuif Cedex, France
severine.guillaume@cnrs.fr, alexis.michaud@cnrs.fr, balthazar.donascimento@cnrs.fr

RÉSUMÉ

La Collection Pangloss, une archive de langues rares, accessibles en ligne, sera présentée en trois parties. Tout d'abord elle sera placée dans le contexte actuel, puis nous ferons une description de son contenu ainsi que de la structure et du format de ses corpus et enfin nous aborderons les différents domaines de recherche pour lesquels elle offre un fort potentiel.

ABSTRACT

The Pangloss Collection, an online archive of rare languages, will be presented in three parts. First of all, it will be placed in the current context, then we will describe its contents, its the structure, and the format of its corpora. Finally we will discuss the various research fields for which it offers a strong potential.

MOTS-CLÉS : corpus oraux, langues peu dotées, archivage, science ouverte, Traitement Automatique des Langues, interdisciplinarité

KEYWORDS: oral corpora, poorly endowed languages, archiving, open science, automatic natural language processing, interdisciplinarity

1 Principes : importance fondamentale d'archives orales pour les sciences du langage

La Collection Pangloss a été créée par le laboratoire de langues et civilisations à tradition orale (LACITO), dans les années 90, dans le prolongement des méthodes classiques d'enquête et d'analyse de langues qui ne possédaient pas de documentation écrite auparavant (Boas 1902; Bouquiaux & Thomas 1971). Les langues en voie de disparition peuvent, à défaut d'être « sauvées », être sauvegardées en les fixant par l'écrit ; à l'ère des technologies numériques, la *documentation et description fondamentale* s'enrichit de composantes audio et vidéo, en plus des trois piliers que constituent dictionnaire, grammaire, et recueil de textes annotés dans les règles de l'art (au sujet des gloses interlinéaires, voir en particulier Lehmann 2004).

Enregistrer les locuteurs à l'aide d'enregistreurs audio ou de caméras permet de recueillir une documentation plus riche qu'une transcription exclusivement écrite. Un enregistrement de la *vive*

voix fournit des informations supplémentaires : la manière de prononcer, l'intonation… véhiculent de nombreuses nuances (Fónagy 1983), avec des moyens qui ne coïncident qu'en partie avec ceux qui peuvent être déployés à l'écrit.
La Collection Pangloss constitue un élément au sein de ce qu'on nomme aujourd'hui volontiers un *écosystème* mondial : elle est l'une des collections de Cocoon, *Collection de corpus oraux numériques*, archive ouverte qui constitue elle-même un des maillons du réseau d'archives ouvertes OLAC (Open Language Archive Community).

2 Caractéristiques techniques

En 2019, la Collection Pangloss regroupe environ 170 langues, avec 1900h d'enregistrements audio/vidéo répartis en 3600 enregistrements dont 1700 possèdent des annotations textuelles (ce qui représente environ 70% du temps d'enregistrement annoté). Des chercheurs et enseignants chercheurs de divers laboratoires ou institutions y déposent aujourd'hui leurs corpus.
Les ressources de la Collection Pangloss sont hébergées par la plateforme Cocoon (Collection de corpus oraux numériques) sur les serveurs de la TGIR Huma-Num. Elles bénéficient des services d'archivage pérenne qui ont été mis en place entre Cocoon, le CNRS et le CINES afin d'assurer un accès et une lisibilité des ressources sur le long terme.
Cocoon étant, de plus, déclaré comme un entrepôt OAI, la Collection Pangloss est moissonnable dans son intégralité par quiconque voudrait avoir accès aux métadonnées descriptives de la collection ainsi que de chacune des ressources qui en fait partie.

3 Perspectives pour la recherche

La quasi-totalité des ressources de la Collection Pangloss est en accès libre (environ 99%), elles sont donc disponibles pour divers usages : découverte, enseignement, recherche. Recherche en linguistique et anthropologie mais aussi, grâce au numérique, recherche dans le traitement automatique des langues.
Il faut noter qu'il est extrêmement rare d'avoir des langues en dangers documentées de manières structurées et en libre accès. En effet, chaque corpus de la collection a pour base des enregistrements audio au format wav ou flac, ou vidéo au format mp4 ou ogg qui sont souvent accompagnés d'annotations textuelles généralement structurées en xml.
La description du contenu des corpus, les métadonnées, suivent les standards actuels (xml, Dublin Core et OLAC pour les spécifications liées à la linguistique) ce qui permet une navigation dans le catalogue de la collection ainsi qu'une détection et une extraction automatisable des données.

Traitement Automatique des Langues : l'intérêt des « petites langues »

De nombreuses perspectives pour le Traitement Automatique des Langues s'ouvrent grâce à la disponibilité de données de « petites langues », qui présentent des scénarios différents de ceux qui se présentent pour les quelques « grandes langues ». De l'ordre de 1% à 10% des langues du monde constituent les principaux enjeux économiques, mais au plan de la recherche en informatique, les défis les plus intéressants ne sont pas nécessairement ceux qui concernent ces « grandes langues ». La mise en ligne en bon ordre de jeux de données de « petites langues » peut donc jouer un rôle moteur dans les progrès de la recherche en Traitement Automatique des Langues.

Le cercle vertueux des collaborations entre linguistes et chercheurs en Traitement Automatique des Langues

Plus il y aura de ressources accessibles en ligne et plus des outils de traitement automatique des langues pourront être mis en place. Plus il y aura d'outils de traitement automatique des langues et plus la description des ressources sera facilitée, et plus on pourra mettre de ressources supplémentaires en ligne. La Collection Pangloss vise à permettre la mise en place de ce cercle

vertueux. Les langues peu dotées se prêtent à des approches multidisciplinaires et innovantes (Besacier 2012). Ainsi, un des corpus de la Collection Pangloss est actuellement utilisé dans un projet de transcription automatique de la parole pour des langues peu dotées. Ce projet a donné naissance à l'outil Persephone qui, à partir de quelques heures d'enregistrements annotés, permet une transcription automatique de nouveaux enregistrements dans cette langue (Adams et al. 2018; Michaud et al. 2018).
Le poster présenté aux Journées scientifiques du Groupement de recherche visera à susciter des échanges permettant de nouer des projets communs autour des ressources, outils et méthodes de la Collection Pangloss.

Références

Adams, Oliver, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird & Alexis Michaud. 2018. Evaluating phonemic transcription of low-resource tonal languages for language documentation. *Proceedings of LREC 2018 (Language Resources and Evaluation Conference)*, 3356–3365. Miyazaki. https://halshs.archives-ouvertes.fr/halshs-01709648.

Besacier, Laurent. 2012. A multi-disciplinary approach for processing under-resourced languages. *Proceedings of International Conference on Asian Language Processing (IALP 2012)*. Hanoi.
Boas, Franz. 1902. *Tsimshian texts* (Bulletin of the Smithsonian Institution. Bureau of American Ethnology 27). Washington: Government Printing Office.

Bouquiaux, Luc & Jacqueline Thomas. 1971. *Enquête et description des langues à tradition orale. Volume I : l'enquête de terrain et l'analyse grammaticale*. 2nd edition 1976. Paris: Société d'études linguistiques et anthropologiques de France.

Fónagy, Ivan. 1983. *La vive voix: essais de psycho-phonétique* (Langages et Sociétés). (Ed.) Louis-Jean Calvet. Paris: Payot.
Lehmann, Christian. 2004. Interlinear morphemic glossing. In Geert Booij, Christian Lehmann, Joachim Mugdan & Stavros Skopeteas (eds.), *Morphologie. Ein internationales Handbuch zur Flexion und Wortbildung. 2. Halbband* (Handbücher Der Sprach- Und Kommunikationswissenschaft 17.2). Berlin: de Gruyter.

Michaud, Alexis, Oliver Adams, Trevor Cohn, Graham Neubig & Séverine Guillaume. 2018. Integrating automatic transcription into the language documentation workflow: experiments with Na data and the Persephone toolkit. *Language Documentation and Conservation* 12. 393–429.

La Construction d'un Corpus Comparative : Une Méthodologie Pour L'étude Des Langues Parlées en Afrique de l'Ouest

Abbie Hantgan[1]
(1) CNRS-LLACAN, 7, rue Guy Môquet - BP 8, 94801 VILLEJUIF (France)
abbie.hantgan-sonko@cnrs.fr

Résumé

Ce article présente un cadre méthodologique pour la construction d'un corpus multilingue comparatif avec les traductions juxtalinéaires composées des neuf langues parlées en Afrique de l'Ouest. Les enregistrements inclus dans le corpus ont été recueillis par le chercheur pendant dix ans au Mali, Burkina Faso, et Sénégal. Un but spécifique de cette recherche est d'examiner quelles sont les similitudes et les différences dans le cadre de contes traditionnels. Bien que les langues inclues dans le corpus ne sont pas apparentées, les contes de plusieurs langues de l'Afrique occidentale ont une tradition dans laquelle le lapin et l'hyène jouent des rôles importants. Ainsi, nous pouvons chercher dans le corpus les caractéristiques intéressantes qui ne seraient pas comparables autrement, pour examiner la grammaire et les marqueurs de discours qui ont des rôles similaires. La Figure 1 donne l'exemple d'une recherche de langues Africaines qui ne sont pas apparentées, et qui ont été documentées et compilées par des auteurs différents, mais qui ont toutes une manière similaire de raconter des contes d'animaux. La structure d'un corpus comparatif en ELAN CorpA (Chanard, 2018), comme celui qui est illustré dans la Figure 1, a l'avantage de pouvoir facilement parcourir de nombreuses annotations avec Regular Expressions pour ELAN (Mosel, 2015), qui sont aussi associées aux enregistrements et meta-data. Une recherche de cet échantillon a révélé que les mots les plus utilisés sont 'lapin' et 'dire', donc, les exemples dans (1) sont de deux langues Dogon documentées par l'auteur et (2) de base de données de langue Goemai, parlée en Nigeria (Hellwig, 2003) et Tabaq, parlée en Sudan (Dimmendaal and Hellwig, 2013). Ces exemples démontrent la comparabilité de contextes similaires, bien que les données soient de sources complètement différentes. La présentation donnera aussi une illustration instructive sur la façon de créer un corpus multilingue utilisant les méthodes linguistiques FLEx et ELAN CorpA, et d'exporter les résultats vers d'autres logiciels.

Abstract

Comparative Corpus Construction : A Methodology for the Study of Spoken West African Languages

This paper presents a methodological framework for constructing an interlinearized multilingual comparative corpus composed of nine spoken West African languages. The recordings contained within the corpus have been collected by the researcher over a ten-year span from Mali, Burkina Faso, and Senegal. A specific aim of the current research is to examine what are the similarities and differences in the story-telling genre. Although the languages illustrated in the corpus are unrelated, many West African languages have a tradition of story-telling where animals such as rabbit and hyena play a crucial role. Thus, we can search the corpus for otherwise in-comparative aspects of interest, to examine grammatical

features and discourse markers within similar roles. Figure 1 gives an example of a search of unrelated languages compiled by different authors from across Africa, all of which have a similar manner of telling animal tales. The format of a comparative corpus in ELAN CorpA (Chanard, 2018) as illustrated in Figure 1 has the advantage of being able to easily search through multiple annotations using Regular Expressions for ELAN (Mosel, 2015), which are then linked to accompanying media and meta-data files. A search of this sample found the two most common words were 'rabbit' and 'say', thus, the examples in (1) from two Dogon languages collected by the author and (2) deposits from the ELAR archive at SOAS from Goemai in Nigeria (Hellwig, 2003) and Tabaq in Sudan (Dimmendaal and Hellwig, 2013). These examples illustrate the comparability of similar contexts even across completely different data sources. The presentation will also provide an instructional illustration of how to create an ELAN corpus using linguistic fieldwork tools FLEx and ELAN CorpA, as well as how to export the corpus files and findings into other readable formats.

Mots-clés : pluralingualism, comparability, sharablity.

Keywords : plurilinguisme, comparabilité, partageable.

Figures and Examples

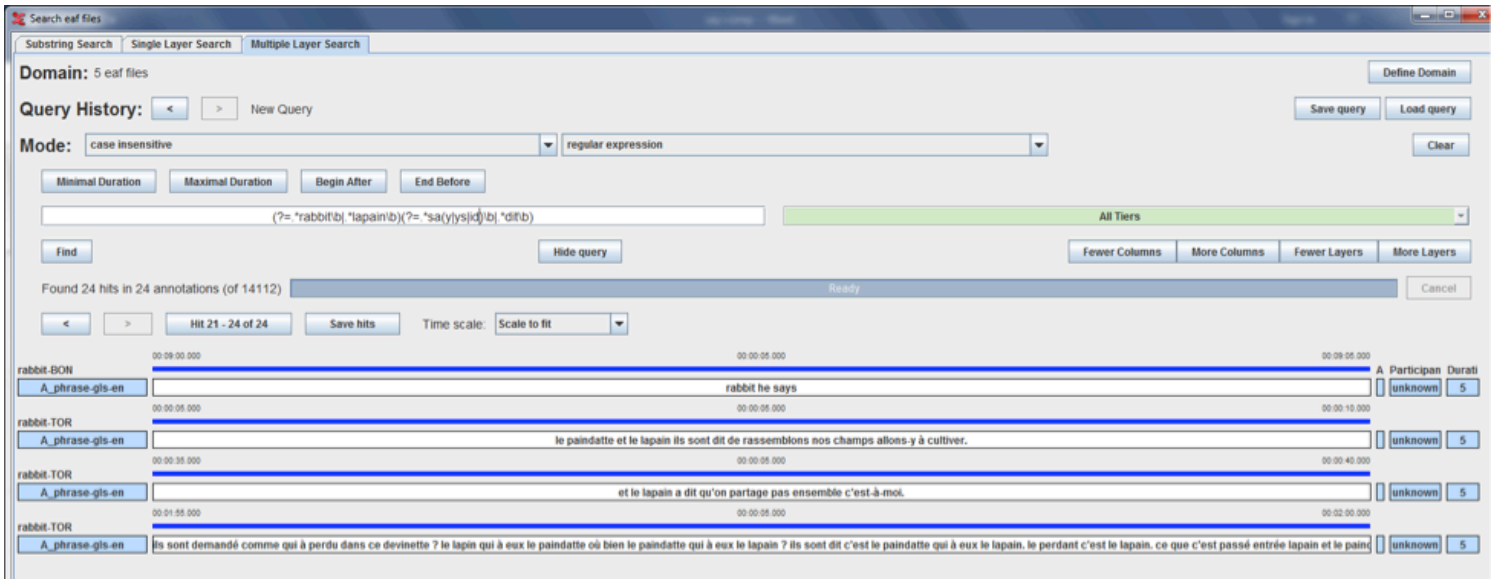


Figure 1 – ELAN Corpus Example

(1) a. ɛ jòn gamu-ru gò wɔ unɔne
COOR lapain partage-NEG dit-IMPV APP 1S
'et le lapain a dit qu'on partage pas ensemble c'est-à-moi'

b. hìyà jɔ́m-ɛ́ mó-dám-á-nj-èè
ok rabbit-NCL.anim 3SG-say-FV-PROG
'ok, rabbit says (to his mother)', 'd'accord, lapin dit (à sa mère)'

(2) a. fûan yi ah to
rabbit say.3S ah ok
'Rabbit says ah ok'
'Lapin dit ah d'accord'

b. aaa budulne k^{w}ala baa
3S rabbit say.PFV COMP
'it was the rabbit that said'
'ce le lapin qui dit'

1 Conférence LIFT 2019 à Orléans

La première édition de la conférence du GDR LIFT se tiendra à Orléans, en plein coeur de la région Centre Val de Loire, célèbre pour ses châteaux, inscrite au patrimoine mondial de l'Unesco et située à 125 km au sud de Paris. Nous accueillerons des participants sur le campus boisé d'Orléans du jeudi 28 novembre 2019 au vendredi 29 novembre 2019.

Le programme comprendra des conférences (conférenciers invités nationaux et internationaux), des sessions poster et des tables rondes.

Les conférenciers invités seront :

- Emily Bender, University of Washington,
- Sabrina Bendjaballah, Université de Nantes,
- Steven Bird, Charles Darwin University & Nawarddeken Academy & University of California Berkeley
- Michel Jacobson, Très Grande Infrastructure de Recherche Huma-Num

1.1 Thème de la conférence

Jusqu'ici, les avancées de la linguistique informatique ont essentiellement débouché sur le développement d'applications : traduction automatique, reconnaissance et synthèse de la parole, détection d'opinion... Or les mêmes avancées (au plan des méthodes symboliques et des modèles d'apprentissage automatique) peuvent également être mises à profit pour faciliter l'analyse linguistique : pour la recherche de généralisations et la création de modèles falsifiables, aussi bien que pour la collecte et la gestion de données. Ainsi, par exemple, les techniques de reconnaissance de la parole peuvent être exploitées pour faciliter le travail des linguistes de terrain ; les techniques de traduction automatique et d'alignement peuvent faciliter la création de lexiques bilingues nécessaires à la documentation d'une langue ; et les algorithmes d'analyse et de génération pour valider des hypothèses syntaxiques et lexicales. La linguistique informatique met à la disposition des linguistes un large éventail de techniques et de ressources qui ouvrent des perspectives nouvelles. L'objectif des Journées scientifiques est de favoriser les interactions entre les trois grands domaines que sont la linguistique informatique, la linguistique formelle et la linguistique de terrain, afin de favoriser le développement de recherches en linguistique qui tirent le meilleur parti des nouvelles technologies.

Les soumissions se feront en vue de proposer un poster à la conférence. L'Appel à communications et la participation aux Journées scientifiques sont ouverts à toutes et tous. À titre non exclusif, les thèmes suivants sont proposés :

- Retours d'expérience concernant l'emploi et/ou le développement d'outils informatiques pour l'analyse linguistique
- Linguistique informatique et Science ouverte : perspectives ouvertes par le partage des données, des outils et des publications
- Modélisation informatique et linguistique formelle (théorie des langages formels, grammaires d'unification, théorie de la preuve...)
- Mise en dialogue des modèles linguistiques et des modèles d'apprentissage automatique (de tous types : approches génératives et discriminantes, approches statistiques

neuronales, approches de type encodeurs-décodeurs...)
— Méthodes non supervisées ou faiblement supervisées pour l'analyse des langues peu dotées, peu écrites ou non documentées
— Réflexions au sujet de l'automatisation des processus d'analyse et de validation

1.2 Format des soumissions

Les propositions de poster feront entre 2 et 4 pages de longueur, au format indiqué par les feuilles de style LaTeX et LibreOffice disponibles sur le site de la conférence.

Références

Chanard, C. (2018). ELAN-CorpA (version 5). logiciel.

Dimmendaal, G. J. and Hellwig, B. (2013). Tamaq tale : lion & rabbit. A Documention of Tabaq, a Hill Nubian language of the Sudan, in its sociolinguistic context. Data deposit at the Endangered Languages Archive at SOAS University of London.

Gardent, C. (2002). Generating minimal definite descriptions. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pages 96–103. Association for Computational Linguistics.

Hellwig, B. (2003). Folktale. Goemai Texts. Data deposit at the Endangered Languages Archive at SOAS University of London.

Keenan, E. and Paperno, D. (2012). Handbook of quantifiers in natural language, volume 90. Springer.

Mosel, U. (2015). Searches with regular expressions in ELAN corpora. draft.

Retours d'expérience concernant le développement de CasEN pour la reconnaissance d'entités nommées

Denis Maurel
Université de Tours, Lifat
denis.maurel@univ-tours.fr

RÉSUMÉ

L'objet de cet affichage est de présenter le système CasEN de reconnaissance d'entités nommées par une cascade de transducteurs (ou plutôt de graphes Unitex) et ses dernières améliorations, suite au projet Istex : lecture et génération de fichiers au format TEI, reconnaissance d'entités en prenant en compte le contexte global, possibilité d'obtenir en sortie des fichiers annotés sous des formats particuliers, en respectant le format d'origine, et de lister, sous une forme adaptable elle aussi, les entités trouvées avec leur nombre d'occurrences. De plus, la plateforme Unitex peut maintenant traiter de manière robuste plusieurs millions de documents et possède un système intégré de scripts.

ABSTRACT

Feedback on the development of CasEN for named entities recognition.

This poster presents the CasEN system of named entities recognition by an Unitex graph cascade and its improvement consecutive to the Istex project: use of TEI files, taken into account of the global context, output in different formats, list of names with their occurrences. Furthermore, Unitex platform now robustly treats some millions of texts and include an integrated system of scripts.

MOTS-CLES : CasEN, entités nommées, cascade, Unitex, projet Istex, TEI, contexte global
KEYWORDS: CasEN, Named entities, cascade, Unitex, Istex project, TEI, global context

Proposition

L'objet de cet affichage est de présenter système CasEN[1] de reconnaissance d'entités nommées par une cascade de transducteurs (ou plutôt de graphes Unitex) et ses dernières améliorations, suite au projet Istex (Maurel et al., 2019).

Le but du projet Istex-Entités nommées était d'apporter une valeur ajoutée à l'interrogation de la base documentaire Istex en autorisant une recherche d'articles via les entités nommées. Cette base documentaire comprend, à ce jour, 23 millions d'articles scientifiques, mis au format TEI.

[1] Disponible sous licence libre à l'URL : http://tln.lifat.univ-tours.fr/Tln_CasEN.html

La reconnaissance d'entités nommées, dont on trouvera un état de l'art un peu ancien dans (Nadeau, Sekine, 2009) et un plus récent, consacré à l'apprentissage dans (Yadav, Bethard, 2018), est aujourd'hui massivement tournée vers l'apprentissage. Cependant l'apprentissage n'est pas un processus entièrement automatique car on omet souvent de mentionner le travail préliminaire de constitution de corpus annotés, avec les difficultés inhérentes à l'annotation et à l'accord inter annotateur (Fort, 2012).

Or les documents de la base Istex proviennent d'époques et d'éditeurs différents et concernent des domaines différents, à la fois dans les "sciences dures" et dans les "sciences de l'homme". Il en résulte une grande hétérogénéité. Et aucun corpus d'apprentissage n'est disponible. Plutôt que de tenter d'en construire un, nous avons donc choisi de ne pas utiliser un système à base d'apprentissage, mais un système symbolique, CasEN (Maurel et al., 2011), composé de cascades (Abney, 1996) (Friburger et Maurel, 2004) de graphes Unitex.

Nous avons adapté CasEN pour qu'il puisse lire et générer des fichiers au format TEI. Nous avons ajouté une reconnaissance d'entités en prenant en compte le contexte global (Wolinski et al., 1995) par généralisation d'étiquetage. Celui-ci ajoute des entrées à son dictionnaire (mais son corpus est homogène) ; contrairement à lui, nous avons choisi comme limite le texte en cours de traitement.

La génération de fichiers par CasEN a aussi été améliorée. Il est facilement possible d'obtenir des fichiers annotés sous des formats particuliers, en respectant le format d'origine. CasEN permet aussi de lister, sous une forme adaptable elle aussi, les entités trouvées avec leur nombre d'occurrences.

Nous avons aussi travaillé sur la plateforme Unitex, afin qu'elle puisse traiter de manière robuste plusieurs millions de documents. Et ajouté un système intégré de scripts pour en simplifier l'utilisation[2].

Nous présenterons aussi les résultats obtenus au cours du projet Istex en terme de précision et de rappel.

Références

Abney S. (1996), Partial Parsing via Finite-State Cascades, *Workshop on Robust Parsing*, 8th *European Summer School in Logic, Language and Information*, Prague, Tchèquie, 8-15.

Fort K. (2012). *Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus*. Thèse de l'université Paris 13.

Friburger N., Maurel D. (2004), Finite-state transducer cascade to extract named entities in texts, *Theoretical Computer Science*, vol. 313, 94-104.

[2] Ces améliorations logicielles d'Unitex ont été réalisées en sous-traitance par Gilles Vollant de la société Ergonotics.

Maurel D., Friburger N., Antoine J.-Y., Eshkol-Taravella I., Nouvel D. (2011). Cascades de transducteurs autour de la reconnaissance des entités nommées[3]. *Traitement automatique des langues*, 52(1):69-96.

Maurel D., Morale E., Thouvenin N., Ringot P., Turri A. (2019). Istex: A Database of Twenty Million Scientific Papers with a Mining Tool Which Uses Named Entities[4]. In Gamallo P., Garcia M., editors. Special Issue "Natural Language Processing and Text Mining". Information 10(5), 178. MDPI.

Nadeau N., Sekine S. A survey of named entity recognition and classification. In Named Entities: Recognition and classification and Use; Sekine, S., Ranchhod, E., Eds.; John Benjamins Publishing Company: Amsterdam, The Netherlands, 2009; pp. 3–28.

Wolinski F., Vichot F., Dillet B. (1995), Automatic processing of Proper Names in Texts, *ACL*.

Yadav V., Bethard S. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 2145–2158

[3] http://www.atala.org/-Varia,81-

[4] https://www.mdpi.com/2078-2489/10/5/178/pdf

Augmentation non Supervisée de Données pour des Langues Peu Dotées non Standardisées

Alice Millour Karën Fort
Sorbonne Université / STIH - EA 4509
28, rue Serpente, 75006 Paris, France
alice.millour@sorbonne-universite.fr, karen.fort@sorbonne-universite.fr

ABSTRACT

Unsupervised Data Augmentation for Less-Resourced Languages with no Standardized Spelling

Non-standardized languages are a challenge to the construction of representative linguistic resources and to the development of efficient natural language processing tools : when spelling is not determined by a consensual norm, a multiplicity of alternative written forms can be encountered for a given word, inducing a large proportion of out-of-vocabulary words.To embrace this diversity, we propose a methodology based on crowdsourcing alternative spellings from which variation rules are automatically extracted. The rules are further used to match out-of-vocabulary words with one of their spelling variants. This virtuous process enables the unsupervised augmentation of multi-variant lexicons without requiring manual rule definition by experts. We apply this multilingual methodology on Alsatian, a French regional language and provide (i) an intrinsic evaluation of the correctness of the obtained variants pairs, (ii) an extrinsic evaluation on a downstream task : part-of-speech tagging. We show that in a low-resource scenario, collecting spelling variants for only 145 words can lead to (i) the generation of 876 additional variant pairs, (ii) a diminution of out-of-vocabulary words improving the tagging performance by 1 to 4%.

MOTS-CLÉS : Langues non standardisées, variation scripturale, myriadisation, annotation en parties du discours.

KEYWORDS: Non-standardized languages, spelling variants, crowdsourcing, part-of-speech tagging.

Résumé en Français

L'article complet décrivant ce travail a été publié en anglais à RANLP 2019 et est accessible ici : `https://hal.archives-ouvertes.fr/hal-02280002`.

Les langues non standardisées, qu'aucun standard n'ait été établi, ou que celui-ci ne soit pas utilisé par les locuteurs, présentent une multiplicité de graphies possibles pour un mot donné. Celles-ci peuvent être causées par l'existence de variantes dialectales au sein de la langue, par les habitudes scripturales variées des différents locuteurs, ou par une accumulation de ces deux phénomènes.

La multiplication des graphies résulte en une forte proportion de mots hors-vocabulaire dans un

contexte où corpus bruts comme annotés sont rares.

Nous présentons ici une méthodologie permettant de prendre en compte cette diversité, basée sur la collecte par myriadisation (cowdsourcing) de graphies alternatives pour un mot donné. Nous extrayons automatiquement des paires de graphies fournies par les locuteurs un ensemble de motifs de variation. Les règles de substitution qui en découlent sont utilisées pour identifier parmi les mot hors-vocabulaires ceux qui sont en réalité une variante graphique d'un mot connu. Ce processus vertueux permet d'accroître le lexique "multi-variantes" initialement fourni par les participants sans nécessiter l'écriture manuelle de règles.

Nous présentons les résultats de cette méthodologie concue comme mutilingue appliquée au cas de l'Alsacien. L'évaluation intrinsèque est réalisée par un expert ayant examiné les paires de variantes découvertes, l'évaluation extrinsèque de la méthode est donnée par l'impact observé sur la tâche d'annotation en parties du discours.

Nous montrons que la collecte de paires de graphies pour 145 mots nous a conduit à la génération automatique de 876 paires de variantes additionnelles, ainsi qu'à une diminution de mots hors vocbulaire améliorant les performances d'annotation de 1 à 4%.

Developing technologies for low resource Uralic languages : the case of North Saami and Komi-Zyrian

Niko Partanen[1], Michael Rießler[2], Thierry Poibeau[3]
(1) University of Helsinki, Finland
(2) University of Eastern Finland
(3) CNRS-École normale supérieure/PSL, U. Sorbonne nouvelle/USPC, France
nikotapiopartanen@gmail.com, m.riessler@gmail.com, thierry.poibeau@ens.fr

ABSTRACT

North Sami and Komi-Zyrian are two Finno-Ugric languages with few linguistic resources available for automatic processing. Based on our experience concerning the development of NLP tools for these languages, we summarize our findings and make some proposals for the application of automatic methods in this context.

RESUME

Mise au point d'outils de TAL pour des langues ouraliennes peu dotées : le cas du same du nord et du komi-zyriène

Le same du nord et le komi sont deux langues finno-ougriennes avec peu de ressources disponibles pour le traitement automatique. Partant de notre expérience concernant la mise au point d'outils efficaces pour ces langues, nous dressons un bilan et proposons différentes pistes pour l'application de méthodes automatiques dans ce contexte.

MOTS-CLES : Ici une liste de mots-clés en français. Times, 10pt.
KEYWORDS: Là, une liste de mots-clefs en anglais. Times, 10pt.

1 Introduction

The Uralic languages form a language family spoken by approximately 25 million people, predominantly in northern and eastern Europe and western Siberia (Wikipedia). Except Finnish, Hungarian and Estonian, all Uralic languages are endangered to some extent. In this context, language technologies can play a major role to document and describe these languages better. While doing so, knowledge about their specificities remain and we can help preserving and teaching them, information technology bringing major benefits to this end.

Originally, language technologies for Uralic have been developed only for written language variants (especially by the Giellatekno center for language technology at the University of Tromsø). In contrast our approach – informed by Computational linguistics and (fieldwork-based) Documentary

linguistics – also includes spoken language data. We have developed technologies for Komi and Saami, based on the most recent advances in natural language processing, applied to a context where resource (especially annotated data) is lacking.

In this poster, we give an overview of the research done (that has been published before in different publications listed below, see more specifically Lim et al., 2018; Partanen et al. 2018a and 2018b) and we detail what are, according to us, the main challenges for language technology in low resource contexts. Our publications are focused on two specific languages but we think similar experiences can be done for a large variety of languages, especially when raw digitized texts are available but other kinds of resources (and specifically annotated data) are lacking.

2 Summary of the research done so far

We have mainly developed our research along three different directions.

1. One is the integration of various language technology in order to get more efficient NLP workflows for language documentation (Gerstenberger et. al. 2017). Compared to traditional NLP workflows for written texts, ours must integrate speech technologies, for speech transcription or signal analysis. As for written texts, processing workflows may go beyond pure NLP, so as to integrate document analysis and OCR.

2. We have also conducted several case studies about dependency parsing in these low resource scenarios (Lim et al., 2018). Dependency parsing is now a relatively mature technology, mainly based on advanced machine learning techniques that require large amounts of annotated data to get accurate results. This is of course a major issue for low resource scenarios, but recent techniques based on multilingual models and language transfer have made it to possible to get working results even in extremely low resource scenarios. The results are of course far from perfect but our aim in the long run is of course to use these methods into the language documentation work. Automatic annotations need to be revised and corrected, but they also make it possible to considerably increase the size of the data produced (which, in turn, makes it possible to train better parsers that will require less manual correction).

3. Third portion of our work has focused more into concrete resource creation, which is illustrated by two Zyrian Komi treebanks (Partanen et al, 2018b) and a large spoken language corpus (Blokland et. al. 2016). This shows that technical advances work hand in hand with the production of resources and help maintain and document endangered languages. This work also aligns closely with observations others have made in relation to this field, namely that even a small amount of annotated data still brings at the moment clear improvements into any multilingual scenarios (Meechan-Maddon et. al. 2019). As our own datasets have grown recently, we are also replicating and extending our earlier experiments.

3 Challenges for future work on under-resourced Uralic languages

Researchers working on endangered languages are facing very practical problems that are not often addressed in the literature because they are practical and not really interesting from a theoretical point of view. NLP researchers are also facing the same kind of problems (to clean the data for example) but probably at a lower scale, and the NLP community is probably better equipped to solve these problems. To take one example, one of the particularities of endangered Uralic languages is that several of them have relatively large written resources, but these resources may not be entirely accessible or available in formats that could be useful in computational linguistics or other linguistic research.

Many resources have already been published as printed text collections and archives both in Europe and Russia contains hundreds of hours of unprocessed recordings. Effective use of modern text and speech recognition methods could help:

- Bootstrapping NLP tools effectively in low-resource scenarios
- Combining the sparse resources in most effective ways
- Working with spoken data: smaller amount of data, higher variation (less standardization) within the data
- Working with endangered languages: different multilingual phenomena are also continuously present, which poses its own challenges.

It means we need to have open source repositories, with corpora in their original formats, but also in more standard ones as far as possible. We also need to have access to different kinds of tools that can be adapted to different tasks and different languages. Maybe more importantly, we need to have more experience reports to give feedback about what works, but also (and above all) what does not, so that newcomers can more effectively analyze their data and produce new annotation and analysis.

At the moment, our more crucial needs are the following:

- Faster and more effective corpus-building (alignment, speech-to-text, tagging) to assist the manual work (creating an annotated linguistic corpus of around 150,000 tokens with traditional methods takes one 3-year-project with two or more full time assistants)
- Work with multimedia: audio and video for fieldwork data, plus additionally images (scanned manuscripts) for legacy data

One should note that these needs are rather general and do not specifically apply to Komi or Saami or even Uralic languages.

Finally, we also need to find the right balance between automation and human expertise, and make people from different backgrounds collaborate. Pure linguists sometimes reject automatic tools because these tools are not perfect. Pure NLP researchers are often not really interested in low

resource languages, because evaluation is difficult when no test set is available, and publication in the NLP community is then nearly impossible (NLP nowadays requires to perform some evaluation and comparison with previous experiments over large datasets, which is, by definition, not possible for low resource languages).

These are real difficulties, but the challenges of our field are important enough so that we can expect these difficulties will not remain forever. After all, NLP is an applied domain, and there is no better context than low resources scenarios to test the robustness and genericity of automatic methods. The need to preserve language diversity is also largely recognized today, and this will help fostering the effort towards the analysis of low resource languages.

Références

Rogier Blokland, Vasiley Chuprov, Maria Fedina, Marina Fedina, Dmitry Levchenko, Niko Partanen & Michael Rießler (2014–2016). Spoken Komi Corpus. Language Bank of Finland Korp version 1.0.0

Ciprian Gerstenberger, Niko Partanen, Michael Rießler, and Joshua Wilbur (2017). Utilizing language technology in the documentation of endangered Uralic languages. Northern European Journal of Language Technology: Special Issue on Uralic Language Technology.

KyungTae Lim, Niko Partanen, Thierry Poibeau (2018). Multilingual Dependency Parsing for Low-Resource Languages : Case Studies on North Saami and Komi-Zyrian. Conference on Language Resource and Evaluation (LREC), Miyazaki, Japan, 2018. < hal-01856178>

Ailsa Meelchan-Maddon & Joakim Nivre (2019). How to Parse Low-Resource Languages: Cross-Lingual Parsing, Target Language Annotation, or Both? Third workshop on Universal Dependencies (UWD 2019), Paris, France.

Niko Partanen, Rogier Blokland, KyungTae Lim, Thierry Poibeau, Michael Rießler (2018a). The First Komi-Zyrian Universal Dependencies Treebanks. Second Workshop on Universal Dependencies (UDW 2018), Brussels, Belgium.

Niko Partanen, KyungTae Lim, Michael Rießler, Thierry Poibeau (2018b). Dependency parsing of code-switching data with cross-lingual feature representations. Fourth International Workshop on Computational Linguistics for Uralic Languages (IWCLUL 2018), Helsinki, Finland.

GREW, un outil au service de l'annotation de corpus et de l'exploitation de corpus annotés

Guy Perrier[1] Bruno Guillaume[1]
(1) LORIA, 54506 Vandœuvre-lès-Nancy cedex, France
guy.perrier@loria.fr, bruno.guillaume@loria.fr

RÉSUMÉ

Nous présentons GREW, un outil dédié à l'annotation de corpus et à l'exploitation de corpus annotés. Fondé sur la réécriture de graphes, il peut être utilisé d'une part pour rechercher des motifs dans des données annotées et d'autre part pour transformer ce type de données.

ABSTRACT

Graph rewriting for corpus annotation and annotated corpus exploitation

We present GREW, a tool dedicated to corpus annotation and the exploitation of annotated corpora. Based on graph rewriting, it can be used to search for patterns in annotated data and to transform this type of data.

MOTS-CLÉS : réécriture de graphes, annotation de corpus.

KEYWORDS: graph rewriting, corpus annotation.

Que ce soit pour la linguistique de corpus ou le TAL fondé sur de l'apprentissage profond ou pas, il est de plus en plus nécessaire de disposer de corpus annotés de qualité et de taille importante, ainsi que d'outils automatiques pour les manipuler. De quelles manipulations, a-t-on le plus souvent besoin ? De pouvoir retrouver un motif dans une annotation, de corriger une annotation, de pouvoir la convertir d'un format dans un autre à un même niveau linguistique, ou encore de produire une annotation à un niveau à partir d'une annotation à un autre niveau. Le développement d'outils automatiques dédiés à ces différentes tâches est facilité par l'inscription de ceux-ci dans un cadre mathématique commun.

1 Le choix du cadre mathématique des graphes

Les graphes s'imposent naturellement pour la représentation des structures sémantiques, ne serait-ce que parce qu'une entité peut être en même temps l'argument de plusieurs prédicats. Les structures syntaxiques, quant à elles, sont en général des arbres, c'est-à-dire des graphes particuliers. Or qui peut le plus, peut le moins. Par ailleurs, il est parfois utile de considérer l'interaction du niveau syntaxique avec un niveau linguistique voisin. Si par exemple, on veut représenter en même temps les relations syntaxiques et les relations d'ordre entre les mots, les structures qui en résultent sont des graphes.

Tout en restant dans le cadre de la syntaxe, plusieurs théories linguistiques (Sgall et al., 1986; Mel'čuk, 1988) ont conçu un niveau intermédiaire (appelé syntaxe profonde) entre la syntaxe (rebaptisée syntaxe de surface) et la sémantique. En syntaxe profonde, il s'agit de représenter uniquement les relations entre mots lexicaux, même quand elles sont indirectes ou implicites. Les structures qui en résultent

sont en général des graphes.

L'outil que nous avons conçu avec Guillaume Bonfante, GREW [1], est fondé sur la réécriture de graphes (Bonfante et al., 2018). Un système de réécriture de graphes est un ensemble de règles qui décrivent des transformations élémentaires et qui sont appliquées successivement pour réaliser une transformation plus globale. Chaque règle est formée d'une partie gauche qui décrit un motif qui doit être remplacé dans un graphe et d'une partie droite qui indique par quoi ce motif doit être remplacé. GREW est utilisé de deux façons en TAL : pour rechercher un motif dans une annotation et pour transformer une annotation.

2 La recherche de motifs dans un corpus annoté

Il est très utile de pouvoir retrouver un motif donné dans un corpus annoté, que ce soit à des fins linguistiques ou pour détecter des erreurs et des incohérences de manière à les corriger.

Si une annotation se présente comme un graphe, et si on considère aussi un motif comme un graphe, le problème revient à apparier deux graphes. Le module de GREW dédié à cette tâche, GREW-MATCH, fournit une syntaxe très simple pour décrire les motifs à rechercher, aussi complexes soient-ils.

Supposons que nous voulions étudier dépendances à distance dans les propositions relatives dans un corpus annoté en syntaxe de dépendances selon le format de Universal Dependencies (UD) [2]. Le chemin de dépendances qui va de l'antécédent du pronom relatif jusqu'au pronom lui-même, en passant par la tête de la relative peut être plus ou moins long.

Le motif suivant, écrit dans la syntaxe de GREW, permet déjà de lister toutes les constructions avec subordonnées relatives :

```
pattern { ANT -[acl:relcl]-> REL_HEAD }
```

Les identifiants `ANT` et `REL_HEAD` sont utilisés pour nommer les nœuds source (l'antécédent d'un pronom relatif) et cible (la tête de la relative) pour pouvoir y faire référence par la suite. L'étiquette `acl:relcl` exprime une dépendance de l'antécédent vers la tête d'une relative.

GREW permet d'ajouter au motif des conditions négatives pour filtrer des solutions que l'on souhaite écarter. Par exemple, enrichissons le motif précédent avec une condition négative :

```
pattern { ANT -[acl:relcl]-> REL_HEAD }
without { PROREL [cat=PRON, Prontype=Rel]; REL_HEAD -> PROREL }
```

La condition négative introduite par le mot-clé `without` permet d'écarter tous les cas où le pronom relatif `PROREL` est rattaché directement à la tête `REL_HEAD` de la relative, et ainsi de lister les cas où le pronom relatif est enchâssé.

Par l'ajout successif de conditions négatives, on peut affiner progressivement la recherche. Par exemple, on peut vouloir écarter de l'étude les cas où le pronom relatif est enchâssé dans un groupe prépositionnel. Il suffit d'ajouter au motif la condition négative suivante.

1. http://grew.fr/
2. https://universaldependencies.org/

```
without {
  REL_HEAD -> PREP; PREP [cat=P];
  PREP -[obj.p]-> PROREL; PROREL [cat=PRO, s=rel];}
```

Cette façon de procéder est très générale et sa mise œuvre est facile car l'outil est disponible en ligne sans installation [3].

3 La transformation d'annotations par réécriture de graphes

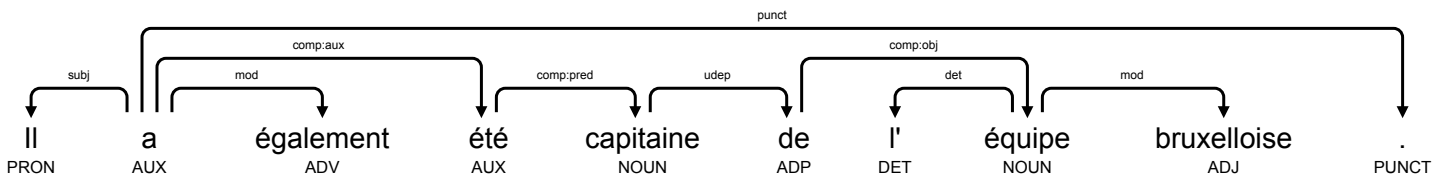
La réécriture de graphes est un prolongement de l'opération d'appariement présentée dans la section précédente. Un système de réécriture de graphes est un ensemble de règles locales de transformations élémentaires de graphe. Une règle est formée d'une partie gauche (qui décrit le motif à rechercher dans le graphe) et d'une partie droite (qui décrit comment modifier le graphe). La difficulté est de préciser comment le résultat de l'application de la règle va être connecté au contexte. Il n'y a pas mathématiquement de façon standard de faire, nous avons donc conçu un modèle de la réécriture de graphes spécifiquement adapté au TAL, où la partie droite des règles se présente comme une suite d'opérations élémentaires réalisant la modification du graphe (ajout ou suppression de nœuds ou d'arcs par exemple).

En pratique, de nombreuses applications de règles sont nécessaires pour effectuer une transformation et il est crucial de contrôler leur enchaînement, ce que permet GREW avec la notion de *stratégie*. Une stratégie est un moyen de décrire l'ordre selon lequel des règles ou des paquets de règles s'enchaînent.

De quelles transformations a-t-on besoin quand on travaille avec des annotations de corpus ? Ce peut être tout d'abord des corrections d'erreurs systématiques d'annotation. Ce peut être aussi produire une annotation à un niveau linguistique donné à partir d'un niveau voisin, par exemple produire une annotation sémantique à partir d'une annotation syntaxique. Ce peut être enfin convertir une annotation d'un format dans un autre format en restant au même niveau linguistique.

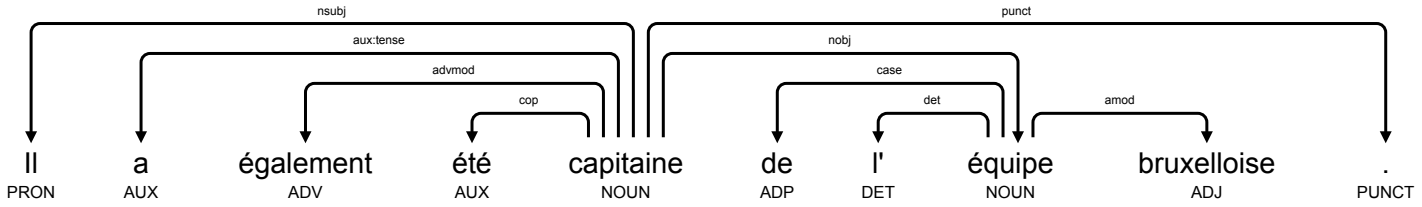
Par exemple, nous avons écrit un système de règles pour convertir une annotation syntaxique du format Universal Dependencies (UD) au format Surface Universal Dependencies (SUD) [4] (Gerdes et al., 2019) et un autre système pour faire la conversion inverse.

Voici un exemple de phrase annoté dans le format SUD.



Voici maintenant la même phrase annotée dans le format UD, l'annotation ayant été obtenu par application de notre système de règles de réécriture.

3. http://match.grew.fr
4. https://surfacesyntacticud.github.io/



Pour obtenir cette annotation, il a fallu appliquer 24 règles. Une différence importante entre les deux formats, que l'on peut remarquer sur l'exemple ci-dessus, est le choix des têtes des dépendances : pour les expressions introduites par des mots fonctionnels (auxiliaires, prépositions, conjonctions), dans SUD, les têtes sont ces mots fonctionnels alors que dans UD, les têtes sont les mots lexicaux.

Voici sous une forme simplifiée la règle qui a effectué le changement de tête des relations *aux :tense* et *cop*.

```
rule rev_head {
  pattern {e:HEAD -[aux:tense|cop]-> AUX;}
  without {HEAD[reversed=y]; AUX[reversed=y] }
  commands {
    HEAD.reversed=y; AUX.reversed=y;
    add_edge e:AUX -> HEAD; del_edge e;
    shift_in HEAD ==> AUX;
    shift_out HEAD =[^unk:fixed]=> AUX}}
```

Elle est formée de deux parties :

— La première partie, avec une clause `pattern` et une clause `without`, décrit le motif à rechercher dans l'annotation. La clause `without` permet de s'assurer que la dépendance n'a pas déjà été retournée. Le fonctionnement de la recherche du motif est identique à ce qu'on a vu dans la section précédente.

— La partie `commands` présente la suite de commandes qui décrivent les modifications souhaitées sur le graphe. Les deux premières commandes consistent à marquer les nœuds `HEAD` et `AUX` du trait `reversed=y` pour indiquer que le retournement est effectué. Les deux commandes suivantes effectuent le retournement. La commande `shift_in` déplace toutes les relations qui arrivent de l'ancienne tête vers la nouvelle et la commande `shift_out` fait de même pour les relations qui partent à l'exception de celles étiquetées `unk:fixed`.

Références

Bonfante, G., Guillaume, B., and Perrier, G. (2018). *Application de la réécriture de graphes au traitement automatique des langues*, volume 1 of *Série Logique, linguistique et informatique*. ISTE editions.

Gerdes, K., Guillaume, B., Kahane, S., and Perrier, G. (2019). Improving Surface-syntactic Universal Dependencies (SUD) : surface-syntactic relations and deep syntactic features. In *TLT 2019, Treebanks and Linguistic Theories, Syntaxfest*, Paris, France.

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

Sgall, P., Hajicová, E., and Panevová, J. (1986). *The meaning of the sentence in its semantic and pragmatic aspects*. Springer Science & Business Media.

Allegro : une plateforme « couteau suisse » pour l'exploitation des ressources textuelles

Étienne Petitjean[1] Christophe Benzitoun[2] Benjamin Husson[1] Sandrine Ollinger[1]
(1) ATILF, Université de Lorraine-CNRS, 54000 Nancy, France
(2) ATILF, CNRS-Université de Lorraine, 54000 Nancy, France
Etienne.Petitjean@atilf.fr, Christophe.Benzitoun@univ-lorraine.fr,
Benjamin.Husson@atilf.fr, Sandrine.Ollinger@atilf.fr

RÉSUMÉ

Nous nous proposons de présenter Allegro, la nouvelle plateforme pour l'exploitation de ressources textuelles développée au sein du laboratoire ATILF, à travers un inventaire rapide de ses applications actuelles et à venir, ainsi que d'une introduction à ses bases techniques. Allegro offre de nombreuses possibilités pour l'indexation et l'interrogation de données structurées, annotées et enrichies de métadonnées.

ABSTRACT

Allegro : A "Swiss knife" platform for exploitation of textual resources

We present Allegro, the new platform for the exploitation of textual resources developed at ATILF. We offer here a quick inventory of its current and future applications, before introducing its technical foundations. Allegro offers many possibilities for indexing and querying structured, annotated and metadata-enriched data.

MOTS-CLÉS : Serveur de données, Ressources linguistiques, Indexation, Interrogation.

KEYWORDS: Data server, Linguistic resources, Indexation, Querying.

1 Introduction

Depuis sa création, le laboratoire ATILF développe, exploite et met à disposition de la communauté scientifique comme du grand public de nombreuses ressources linguistiques (Bernard et al., 2002). Le moteur de recherche Stella (Dendien, 1991), ayant permis d'exploiter ces ressources jusqu'à récemment, a fait son temps. Conçu à la fin des années 80 avec les contraintes et les limitations de l'époque, il n'est pas en mesure de répondre aux nouveaux enjeux informatiques et manque d'adaptabilité. C'est pourquoi il a dû être abandonné. L'idée a donc germé de développer un nouveau moteur, qui serait évolutif et plus facilement maintenable dans le temps : Allegro [1]. Initialement pensé comme un nouveau concordancier pour la base de données textuelles Frantext (Montémont, 2014), Allegro a intégré les besoins de différents projets portés par l'ATILF au fur et à mesure de son développement.

On pourrait, à juste titre, se demander quel est l'intérêt de développer un nouvel outil pour l'exploitation des corpus, alors que certains instruments, tels que Corpus Workbench (Evert and Hardie, 2011),

1. Étienne Petitjean est à l'origine de la majeure partie des développements d'Allegro.

offrent déjà des performances tout à fait appréciables. Si nous nous sommes largement inspirés de la syntaxe de requêtes CQL, étant donné que celle-ci est transparente et largement utilisée dans la communauté des linguistes, Allegro permet une variété d'applications plus étendues. Nous avons également développé nos propres algorithmes de recherche pour optimiser les temps de réponse lors de la recherche de formes dans un lexique ou de l'exécution de requêtes sur corpus. Pour l'instant, Allegro tourne uniquement sur les serveurs de l'ATILF, mais les sources seront disponibles et publiques très bientôt. L'objectif, à moyen terme, est de proposer à la communauté scientifique un instrument gratuit, dérivé de Frantext, qui permettrait aux utilisateurs de définir leurs propres dépôts avec leurs fichiers et leurs métadonnées. Les chercheurs seraient ainsi autonomes pour créer et interroger leur corpus, choisir de les partager ou les garder privés. Un tel service pourrait trouver sa place sur la plateforme Ortolang.

Dans notre communication, nous présenterons les applications actuelles d'Allegro [2] et introduirons quelques aspects techniques de ce nouveau serveur de données pour l'exploitation de ressources textuelles et lexicales. Nous montrerons à travers ces quatre projets qu'Allegro se prête à des visées scientifiques variées, par le traitement et la visualisation de données de natures très différentes. Il offre ainsi la possibilité d'imaginer de nombreuses applications. En réduisant les efforts de transformation des données et en augmentant l'autonomie entre le cœur des applications Web et leurs interfaces, il permettra à l'avenir de concentrer nos efforts de développement informatique sur l'implémentation de nouvelles fonctionnalités, répondant au mieux aux besoins de la recherche en linguistique, en accompagnement de son évolution.

2 Ressources utilisant Allegro

Les fonctionnalités d'Allegro se sont construites à partir des besoins de cinq projets majeurs, dont quatre sont d'ores et déjà implémentés. Dans cette section, nous ferons le point sur la base Frantext et sa refonte récente, avant de présenter la base Aliento et les projets exploitant respectivement les sources du Dictionnaire de l'Académie Française et du Französiches Etymologisches Wörterbuch (FEW). Enfin, nous terminerons par évoquer le projet de refonte de portail lexical actuellement disponible sur le site du Centre National de Ressources Textuelles et Lexicales (CNRTL).

Frantext. Rendue disponible sur abonnement à travers le logiciel Stella en 1992, la base de données Frantext regroupe aujourd'hui 5 415 ouvrages, en majorité littéraire, pour plus de 250 millions de mots écrits entre 1125 et nos jours. Elle a fait l'objet de différentes évolutions au cours de ces dernières années, comme son annotation complète en parties du discours. Depuis avril 2018, c'est l'ensemble de l'interface d'interrogation et des fonctionnalités qui ont fait peau neuve en se basant désormais entièrement sur Allegro. Pour ce faire, il a fallu concilier les habitudes des utilisateurs et le besoin de nouveautés permettant des exploitations plus poussées.

Aliento. Depuis 2007, l'ATILF est partenaire du projet Aliento, qui vise la mise au point d'une méthodologie permettant l'étude de l'évolution à travers les textes, les époques et les langues, des énoncés proverbiaux et, plus spécifiquement, des énoncés sapientiels brefs (ESB). La base Aliento (Bornes-Varol et al., 2018) décrit aujourd'hui plus de dix mille ESB, répartis entre vingt-deux textes médiévaux écrits en arabe, hébreu, espagnol, catalan et latin et maintenus au format XML-TEI.

2. URL : `https://www.frantext.fr`, `https://www.aliento.eu`, `https://academie.atilf.fr`, `https://few-webapp.atilf.fr/`

Dictionnaire de l'Académie française. L'ATILF assure la conversion au format XML-TEI des articles du Dictionnaire de l'Académie française (pour les éditions 4, 7 et 9). Dans ce cadre, nous avons également réalisé une interface d'interrogation basée sur Allegro. Il s'agit d'un outil d'aide pour les lexicographes en charge de la rédaction des articles. Il permet de rechercher dans des sous-parties spécifiques des articles et de mettre en exergue des éléments de structure.

FEW rétroconverti. Le dictionnaire étymologique et historique du galloroman Französisches Etymologisches Wörterbuch (FEW) (Carles et al., 2019) est en cours de rétroconversion par traitement automatique à l'ATILF. Les trois volumes déjà rétroconvertis forment un corpus d'environ 3 000 articles. Les spécificités typographiques et structurelles de cet ouvrage le rendent délicat à manipuler et à exploiter à l'aide des instruments existants. Allegro nous a permis de répondre à ces besoins spécifiques et joue aujourd'hui un rôle dans la chaîne même de traitement de rétroconversion, en permettant d'être au plus proche de la version XML de la ressource interrogée et d'en diagnostiquer aisément les imperfections.

Portail lexical. Le portail lexical du CNRTL regroupe un ensemble de ressources linguistiques et d'outils sous la forme d'une interface de consultation simple et conviviale. Il répond à plus de 700 000 requêtes par jour. Développé en 2006, il utilise les technologies Web de l'époque (XHTML, PHP). Le poids de l'âge commence à se faire sentir, particulièrement au niveau des interfaces graphiques. Les ressources utilisées subissent un traitement spécial pour modifier leur structure afin de les insérer dans une base de données relationnelle. Chaque ressource fait l'objet d'un traitement coûteux en termes de développement et de temps. Nous envisageons une refonte complète de l'application en 2020. Elle s'accompagnera d'une mise à jour technique majeure en utilisant exclusivement Allegro pour stocker et exploiter toutes les ressources visibles sur le portail.

3 Quelques considérations techniques

Allegro est constitué de trois composants logiciels distincts : un indexeur, un environnement d'exécution et un serveur. L'indexeur prend en entrée les données et les métadonnées, il les restructure et produit un format de sortie optimisé permettant de faire des recherches efficaces aussi bien sur les données que sur leur structure. L'environnement d'exécution est le cœur du système. Il permet de définir un corpus, d'effectuer des requêtes et de récupérer les résultats dans le format de sortie choisi. Le serveur est le composant de plus haut niveau. Il encapsule l'indexeur et l'environnement d'exécution pour donner accès à toutes leurs fonctionnalités depuis un serveur Web. Ce composant gère également les autorisations d'accès aux ressources. Allegro offre une interface REST permettant de l'interroger et de l'administrer. Il est ainsi facile de l'utiliser à partir de n'importe quel langage, comme nous le faisons actuellement en TypeScript et en Go.

Les ressources au cœur des projets de la section précédente peuvent toutes être considérées comme des corpus textuels semi-structurés, de taille importante, que nous souhaitons partager avec la communauté scientifique en les accompagnant d'instruments d'exploration. La majeure partie d'entre elles sont en cours d'évolution, ce qui implique des mises à jour régulières. Bien que ces ressources disposent chacune de leurs propres modalités d'interrogation et d'affichage, elles requièrent toutes la réalisation de requêtes portant à la fois sur le contenu textuel et la structure des documents qu'elles regroupent, auxquels viennent parfois s'ajouter différentes couches d'annotation du contenu textuel.

Pour répondre à ces besoins récurrents, Allegro a été conceptualisé et développé de manière à

simplifier la mise à disposition des ressources. Il offre la possibilité de les exploiter sans transformation préalable, quel que soit leur format (texte brut, XML, CSV), pour peu qu'elles soient encodées en UTF-8. Nous limitons ainsi la multiplication de versions parallèles des ressources et nous épargnons l'écriture de chaînes de traitements spécifiques. L'ensemble des métadonnées associées aux ressources doit pour sa part être fourni dans un format JSON. Il n'y a aucun schéma de base à respecter pour chaque entrée. Le format est entièrement libre et nous pouvons donc ajouter n'importe quel type de métadonnées (chaîne de caractère, entier, flottant, booléen, null). Ces métadonnées sont ensuite directement interrogeables. Dans le cas de Frantext, chaque œuvre est ainsi associée à son année de publication, son auteur, son genre, etc., qui deviennent autant de facettes dans son interface d'interrogation utilisées pour constituer les corpus de travail.

La taille des ressources et les temps d'exécution ont également été pris en considération. Allegro permet d'exploiter des corpus textuels comportant autant de couches d'annotation que souhaité et la taille de ces corpus n'est limitée que par la RAM disponible. L'indexation de l'intégralité de la base Frantext se réalise en une minute environ.

4 Conlusion

Le logiciel Allegro, développé au départ comme un simple concordancier, a évolué pour devenir une plateforme complète destinée à simplifier la mise en ligne et l'exploitation des ressources textuelles et lexicales. Déjà dotée de riches fonctionnalités dans sa version actuelle, la plateforme, grâce à son architecture modulaire, permettra de nombreuses évolutions (nouveaux formats en entrée, nouveaux types de requêtes, etc.) lorsque de nouveaux besoins apparaîtront.

Références

Bernard, P., Lecomte, J., Dendien, J., and Pierrel, J.-M. (2002). Un ensemble de ressources informatisées et intégrées pour l'étude du fançais : FRANTEXT, TLFi, Dictionnaire de l'Académie et logiciel Stella, présentation et apprentissage de leur exploitation. In *Actes. 9ème Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN). Nancy, palais des congrès. 24-27 juin 2002*, volume 2, pages 3–36.

Bornes-Varol, M.-C., Husson, B., and Ortola, M.-S. (2018). La base de données Aliento : Bilan. *Aliento : échanges sapientiels en Méditerranée* , 10 :5–12.

Carles, H., Dallas, M., Glessgen, M., and Thibault, A. (2019). *Französiches Etymologisches Wörterbuch, Guide d'utilisation*. Bibliothèque de Linguistique Romane, Hors série 5. Éditions de linguistique et de philologie.

Dendien, J. (1991). Access to information in a textual database : access functions and optimal indexes. In *Research in Humanities Computing, Papers from the 1989 ACH-ALLC Conference, Oxford : Clarendon Press*.

Evert, S. and Hardie, A. (2011). Twenty-first century corpus workbench : Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference*. University of Birmingham.

Montémont, V. (2014). Frantext, une base de données pour la recherche. In *Corpus de textes écrits et oraux : quels usages pour la recherche*, Mons, Belgium. Michel Berré.

Un corpus du kriol à l'épreuve du TAL pour l'étude de la variation

Jean-Louis Rougé Flora Badin
LLL UMR7270, 10 rue de Tours, 45 065 Orléans Cedex 2, France
jean-louis.rouge@univ-orleans.fr, flora.badin@univ-orleans.fr

RÉSUMÉ

Les créoles sont nés du contact entre différentes langues dont une, généralement européenne, qui a apporté le lexique. Le créole est souvent resté au contact de sa langue lexificatrice ce qui est source d'une grande variation. Le kriol parlé en Guinée Bissau et en Casamance, est resté en contact non seulement avec le portugais, langue lexificatrice, mais aussi avec les langues africaines qui ont participé à son émergence. Cette situation engendre une multiplicité de façons de parler cette langue. Afin d'analyser la variation et de trouver les outils pour catégoriser les façons de parler, un corpus a été élaboré. Á la pratique, il s'avère que le traitement de ce corpus requiert un dialogue entre linguistes specialistes du kriol et specialistes du Traitement Automatique des Langues (TAL).

ABSTRACT

A kriol corpus, processed within a NLP framework to study variation

Creoles were born from the contact between different languages, one of which was usually European, and provided the lexicon. Creole has often remained in contact with its lexifying language, which is a source of great variation. The kriol spoken in Guinea Bissau and Casamance, stayed in contact not only with Portuguese, its lexifying language, but also with the African languages that participated in its emergence. This situation generates multiple ways of speaking this language. In order to analyze the variation and find tools to categorize these various forms of speech, a corpus has been created. In practice, processing this corpus requires a collaboration between linguists specialized in kriol and specialists in Natural Language Processing (NLP).

MOTS-CLÉS : Variation,créole, contact de langues, corpus, TAL.

KEYWORDS: Variation, creole, langage contact, corpora, NLP.

1 Créoles et variations

1.1 La question des créoles

S'il n'existe pas de définition des créoles recueillant l'agrément de tous les linguistes, on peut cependant donner une ou deux particularités : ce sont des langues nées du contact entre deux (ou plus) autres langues. Une de ces langues fournit l'essentiel du lexique. Ainsi lorsque l'on parle de créole français, portugais ou anglais on indique simplement que le lexique du créole provient majoritairement de ces langues. L'histoire des pays où les créoles sont parlés fait que très souvent le créole et sa langue lexificatrice sont restés en contact. Le créole est généralement, la langue

du quotidien, la langue la plus employée mais ne bénéficie d'aucun statut. En revanche sa langue lexificatrice est la langue officielle, langue de scolarisation et de la promotion sociale. Au point de vue linguistique, cela se traduit par une grande variation entre des façons de parler qui se rapprochent ou s'éloignent de la langue lexificatrice. Ces situations sont généralement décrites comme un continuum de variétés – basilecte, mesolecte, acrolecte. La réalité est souvent moins simple. Le cas du kriol, créole portugais parlé en Guinée-Bissau et en Casamance (sud du Sénégal) avec ses spécificités est un exemple d'une situation sociolinguistique particulièrement complexe.

1.2 Le kriol

Le kriol est un des créoles portugais d'Afrique. A la différence du créole cap-verdien et des créoles portugais du Golfe de Guinée (forro, angolar, lung'Ie et fa d'Ambu), il s'agit d'un créole continental, ce qui a des conséquences importances. En effet, il s'est développé et évolue non seulement au contact du portugais, mais aussi des différentes langues africaines qui ont participé à son émergence, ce qui contribue amplement a renforcé la variabilité des pratiques et des productions linguistiques. Ces dernières années des éléments externes ont renforcé cette tendance. Le kriol qui jusque-là en Guinée, était essentiellement parlé dans les villes est devenu en l'espace de quarante ans la langue véhiculaire de la quasi-totalité de la population. Si les nouveaux locuteurs du kriol parlent aussi différentes langues régionales, la massification de l'enseignement en portugais ainsi que le développement de nouveaux media et de nouvelles technologies de la communication ont renforcé la présence du portugais. Ainsi, les façons de parler le kriol se diversifient chaque fois plus, posant souvent la question de savoir ce qui appartient au domaine du kriol ou à celui du portugais. Dans ce contexte une analyse en termes de continuum de variétés plus ou moins bien définies (Couto and Embaló, 2011) est illusoire.

Dans le même temps en Casamance, devenue francophones depuis la cession de cette région à la France en 1886, le kriol s'est éloigné du portugais et a fortement subi les influences du français et du wolof (Rougé, 2013) (Nunez, 2015). Le développement récent de nouvelles dynamiques transfrontalières a rétabli les contacts rompus entre le kriol tel que parlé en Guinée et en Casamance et, ici aussi, de nouvelles façons de parler cette langue se sont multipliées.

Afin d'analyser cette situation linguistique, de catégoriser les différentes productions linguistiques, nous avons élaboré un corpus.

2 Un corpus kriol

Pour élaborer ce corpus, nous avons réalisé, à ce jour, 47 enregistrements (effectués par J.-L. Rougé) d'une durée allant de 10 minutes à plus d'une heure, pour un total de 27 heures environ. Ces enregistrements ont été effectués avec un enregistreur numérique Marantz sur plusieurs terrains : en Casamance, à Bissau et à Orléans (2 enregistrements). Comme il s'agit d'enregistrements de discours spontané, ils ont tous les défauts des enregistrements de terrain : bruits de fond, interruptions, etc. Il s'agit d'entretien mené en kriol avec des témoins adultes des deux sexes, d'origines sociales et de niveaux académiques différents (i.e. aussi bien analphabètes que titulaires de diplômes du secondaire, voire du supérieur). Une attention particulière a été apportée au répertoire linguistique de chacun (Kriol langue maternelle ou seconde, exposition au portugais, aux langues africaines, au français…) Au fur et à mesure du traitement des enregistrements (transcriptions, gloses, traduction…) des

questions relatives en particulier à la variation entre les productions, mais aussi, et peut-être surtout à l'intérieure d'une même production se sont fait sentir et pour lesquelles un dialogue avec des spécialistes du TAL pourrait enrichir les pratiques des uns et des autres.

3 Questions

Ce grand corpus oral de kriol, transcrit orthographiquement pour certains enregistrements, est une ressource précieuse pour les spécialistes du TAL souhaitant travailler sur les langues peu dotées. A l'heure actuelle nous menons plusieurs pistes de réflexion pour permettre à ce corpus de voir le jour et valoriser cette langue auprès de ses locuteurs. Quels outils innovants ou quelles méthodes d'apprentissage automatique (par exemple) pourraient nous venir en aide concernant la possibilité de transcrire automatiquement le reste du corpus orthographiquement ou phonétiquement ? La technique de collecte proposée par (Blachon et al., 2016) a retenu notre attention mais semble cependant convenir uniquement à la nouvelle collecte et non pour un corpus déjà existant. En ce qui concerne le contact des langues, nous envisageons de travailler sur la prosodie via un découpage automatique aux mots à l'aide du logiciel Jtrans (Cerisara et al., 2009) et comparer si la courbe mélodique de chacun d'eux diffère qu'il s'agisse d'un emprunt ou non à l'aide du logiciel SLAM (Obin et al., 2014).

Blachon, D., Gauthier, E., Besacier, L., Kouarata, G.-N., Adda-Decker, M., and Rialland, A. (2016). Collecte de parole pour l'étude des langues peu dotées ou en danger avec l'application mobile lig-aikuma.

Cerisara, C., Mella, O., and Fohr, D. (2009). Jtrans: an open-source software for semi-automatic text-to-speech alignment. In *Tenth Annual Conference of the International Speech Communication Association*.

Couto, H. H. d. and Embaló, F. (2011). Literatura, língua e cultura na guiné-bissau. *PAPIA: Revista Brasileira de Estudos Crioulos e Similares*, 20:11–253.

Nunez, J. J. F. (2015). *L'alternance entre créole afro-portugais de Casamance, français et wolof au Sénégal: une contribution trilingue à l'étude du contact de langues*. PhD thesis.

Obin, N., Beliao, J., Veaux, C., and Lacheret, A. (2014). Slam: Automatic stylization and labelling of speech melody.

Rougé, J.-L. (2013). Créole de casamance. émergence de nouvelles variétés. *Travaux-Cercle linguistique d'Aix-en-Provence*, 24:201–212.

Une métagrammaire TAG pour le créole guadeloupéen

Emmanuel Schang
Université d'Orléans & LLL (UMR7270), 10 Rue de Tours - BP 46527, 45065 Orléans cedex 2 (FRANCE)
emmanuel.schang@univ-orleans.fr

RÉSUMÉ

Ce travail présente une métagrammaire TAG au format XMG pour le créole guadeloupéen. Il en décrit les traits principaux reposant sur des analyses linguistiques. Dans un premier temps, je présente le formalisme utilisé puis je donne les principales caractéristiques de cette grammaire.

ABSTRACT

A TAG Metagrammar for Guadeloupean Creole

This work describes a TAG metagrammar for Guadeloupean Creole. It presents the key features of this grammar.

MOTS-CLÉS : Métagrammaire, grammaire d'arbres adjoints, XMG, créoles, gwadloupéyen.

KEYWORDS: Metagrammar, Tree-Adjoining Grammar, XMG, Creole, Guadeloupean French Creole.

1 Introduction

Je présente ici une grammaire électronique du créole de Guadeloupe. Cette grammaire au formalisme TAG (Tree-Adjoining Grammar, cf. (Joshi et al., 1975)(Joshi, 2012)) utilise le formalisme XMG (eXtensible Metagrammar, cf. (Crabbé et al., 2013)(Petitjean, 2014)). Ce travail répond à plusieurs objectifs. Il s'agit de créer une grammaire à base de connaissances (par opposition à une grammaire construite à partir d'un corpus annoté) d'une langue assez peu décrite, pour laquelle l'outillage linguistique est faible. Cela vise à accroître le niveau de connaissances sur cette langue en proposant un large éventail de structures grammaticales décrivant le fonctionnement de cete langue. C'est aussi un moyen de valoriser une langue de France trop souvent stigmatisée (de nombreux locuteurs parlent encore de 'patois' ou de 'mauvais français' en évoquant le créole). Les outils choisis permettent un travail ouvert et cumulatif. Dans un premier temps, je décris le formalisme choisi. Je passe ensuite en revue les principaux points de cette grammaire.

2 TAG et XMG

Les grammaires TAG [1] font partie des formalismes utilisés pour la confection de grammaires électroniques dans différentes langues (v. (Abeillé, 2002) pour le français, (Kallmeyer et al., 2008a) pour

1. v. https://en.wikipedia.org/wiki/Tree-adjoining_grammar pour une description sommaire du formalisme.

FIGURE 1: Constitution de l'arbre dérivé de *Jan manjé anpil* ('Jean a beaucoup mangé')

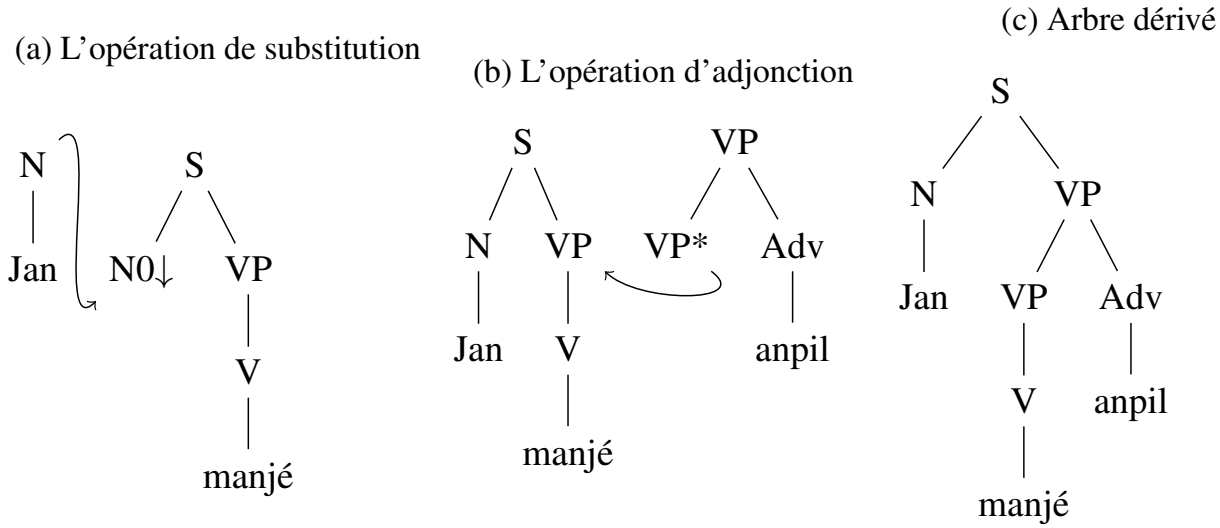
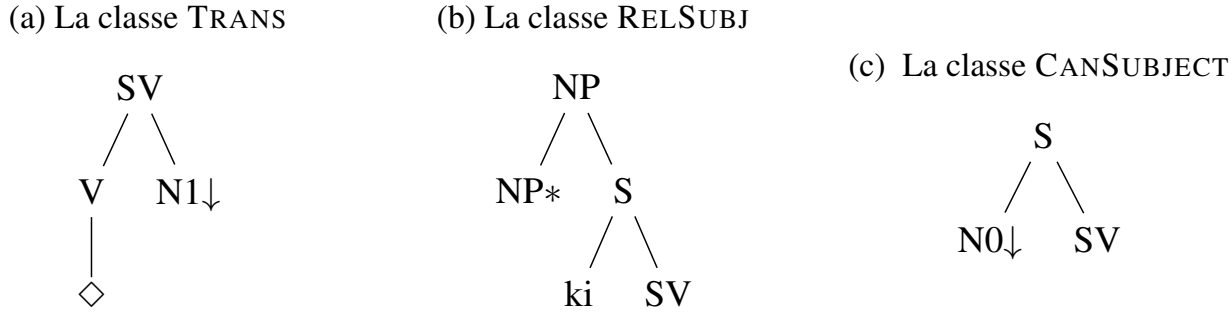


FIGURE 2: Représentation des classes XMG TRANS, RELSUBJ et CANSUBJ



l'allemand et (Han et al., 2000) pour le coréen et (Fraj et al., 2008) pour l'arabe notamment). J'utilise ici une grammaire TAG lexicalisée (dite LTAG) qui n'admet que des arbres ayant au moins une ancre lexicale (un item lexical en noeud feuille). Le fonctionnement des TAG repose sur deux opérations : la *substitution* et l'*adjonction* (adjoining) représentées par un exemple en Fig. 1a et Fig. 1b. Des principes linguistiques guident la formation de ces arbres (cf. (Abeillé, 2002) notamment).

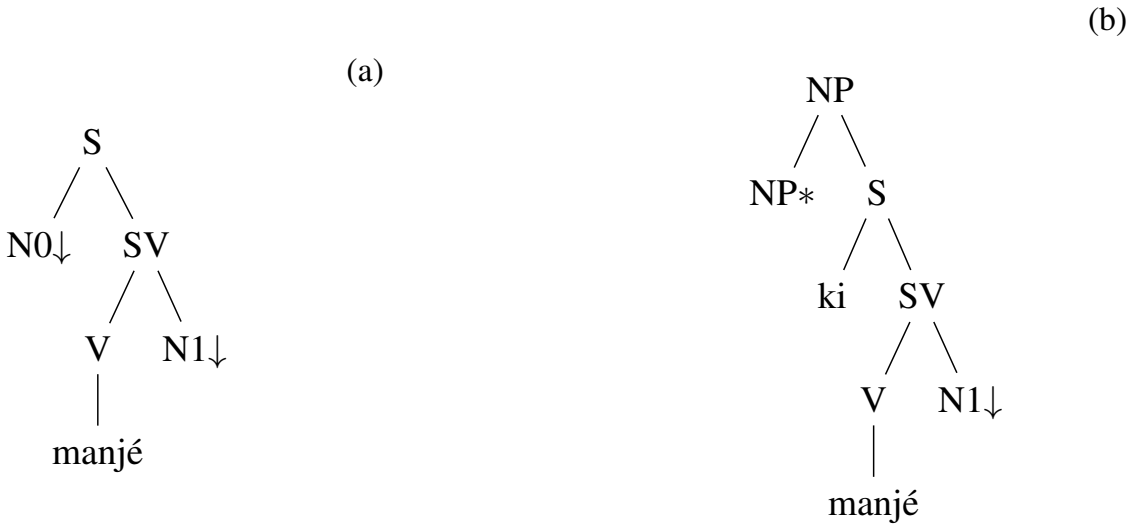
Le formalisme utilisé ici pour la description (et la constitution) des arbres élémentaires TAG de cette grammaire est XMG, grâce à l'outil XMG2 (v. (Petitjean, 2014)). XMG2 permet de décrire une grammmaire et de compiler la description pour obtenir une ressource au format XML lisible par un parser syntaxique (ici TuLiPa (Kallmeyer et al., 2008b)). Concrètement, le linguiste décrit des abstractions sur les arbres élémentaires TAG, sous la forme de *Classes*, qu'il combine par un jeu de disjonctions et conjonctions.

Ainsi, à partir des 3 classes représentées en Fig. 2, on peut obtenir les deux arbres élémentaires licites de la Fig 3.

On peut donc représenter les deux arbres bien formés en Fig. 3 sous la forme d'une conjonction :

— TRANS et CANSUBJ forment l'arbre en 3a,
— TRANS et RELSUBJ forment l'arbre en 3b.

FIGURE 3: Arbres TAG pour N0 MANJÉ N1 et N KI MANJÉ N1



3 Points clés

J'expose ici rapidement les points clés de cette grammaire :

- les prédicats verbaux et non verbaux (prédicats adjectivaux, locatifs, etc.) sont pris en compte (voir aussi (Vaillant, 2008)) ;
- les marqueurs préverbaux (Temps, Mode, Aspect, Négation) figurent en co-ancre des prédicats (verbaux ou non-verbaux) et sont mis en facteur pour les différents types de prédicats. Ils sont considérés comme des projections fonctionnelles sémantiquement dépendantes du prédicat ;
- les projections fonctionnelles du nom (défini spécifique, pluriel, démonstratif, génitif) sont traitées comme co-ancres du nom ;
- le niveau de projection (trait PROJ) permet de gérer les contraintes de placement des éléments fonctionnels.

L'utilisation d'une métagrammaire permet l'implémentation aisée des points énoncés ci-dessus. Par exemple, les classes TMA en Fig. 4 se combinent au sein des arbres élémentaires en fonction du trait PROJ. Cela permet d'autoriser la combinaison *Jan té ka manjé* 'Jean était en train de manger' et de rejeter * *Jan ka té manjé*.

FIGURE 4: Exemples de classes de marqueurs de Temps, Mode et Aspect

4 Perspectives

A l'heure actuelle, la combinaison de 47 classes permet d'engendrer 567 formes d'arbres distinctes. Cette grammaire est conçue à partir de connaissances : (Bernabé, 1983) notamment, ainsi que des

exemples recueillis sur le terrain ou avec l'aide d'étudiants. Ces connaissances sont étendues grâce à des collègues de Guadeloupe qui contribuent à l'accroissement du nombre de formes de phrases traitées [2]. Elle est disponible gratuitement et librement sur GitHub (`https://github.com/eschang/xmg_GC_metagrammar`) pour tous les usages. En particulier, elle peut être étendue dans le domaine sémantique ou servir de support pour la confection de ressources sur des langues proches.

Références

Abeillé, A. (2002). *Une Grammaire électronique du Français*. CNRS Editions, Paris.

Bernabé, J. (1983). *Fondal-natal*. l'Harmattan, Paris.

Crabbé, B., Duchier, D., Gardent, C., Roux, J. L., and Parmentier, Y. (2013). XMG : eXtensible MetaGrammar. *Computational Linguistics*, 39(3) :1–66.

Fraj, F. B., Zribi, C. B. O., and Ahmed, M. B. (2008). Arabtag : a tree adjoining grammar for arabic syntactic structures. In *Proceedings of the international arab conference on information technology*.

Han, C.-h., Yoon, J., Kim, N., and Palmer, M. (2000). A feature-based lexicalized tree adjoining grammar for korean. *IRCS Technical Reports Series*, page 35.

Joshi, A. K. (2012). Tree-adjoining grammars. *Oxford Handbooks Online*.

Joshi, A. K., Levy, L. S., and Takahashi, M. (1975). Tree Adjunct Grammars. *Journal of Computer and System Science*, 10 :136–163.

Kallmeyer, L., Lichte, T., Maier, W., Parmentier, Y., and Dellert, J. (2008a). Developing a tt-mctag for german with an rcg-based parser.

Kallmeyer, L., Lichte, T., Maier, W., Parmentier, Y., Dellert, J., and Evang, K. (2008b). Tulipa : Towards a multi-formalism parsing environment for grammar engineering. In *Proceedings of the Workshop on Grammar Engineering Across Frameworks*, pages 1–8. Association for Computational Linguistics.

Petitjean, S. (2014). *Génération modulaire de grammaires formelles*. PhD thesis, Université d'Orléans.

Vaillant, P. (2008). Grammaires factorisées pour des dialectes apparentés. In Béchet, F., editor, *TALN 2008 : Actes de la 15ème conférence annuelle sur le Traitement Automatique des Langues Naturelles*, pages p. 159–168, Avignon, France. ATALA (Association pour le Traitement Automatique des Langues). 10 pages. Actes de la 15eme conference annuelle sur le Traitement Automatique des Langues Naturelles (TALN 2008), Avignon, France, 9-13 juin 2008.

2. Je remercie Juliette Sainton notamment pour sa contribution.

Interdisciplinary Approach to the Study of Pragmatic Markers in Everyday Spoken Discourse

Natalia V. Bogdanova-Beglarian[1], Olga V. Blinova[1], Tatiana Y. Sherstinova[1, 2]
(1) St. Petersburg State University, Universitateskaya emb. 11, 199034, St. Petersburg Russia
(2) National Research University Higher School of Economics, 190068, St. Petersburg, Russia
{n.bogdanova, o.blinova, t.sherstinova}@spbu.ru

L'approche interdisciplinaire de l'étude des marqueurs pragmatiques dans le discours de tous les jours

ABSTRACT

Interdisciplinary Approach to the Study of Pragmatic Markers in Everyday Spoken Discourse.

Pragmatic markers (PMs) are indispensable elements of spoken discourse in any language. They are speech elements, having major influence on a pragmatic aspect of spoken discourse and being practically devoid of their own referential meaning. In spite of PMs wide circulation, they are very poorly studied. The current research demonstrates an interdisciplinary approach to study of PMs based on two representative speech corpora – ORD corpus of Russian Everyday Speech known as "One-Day-of-Speech"-corpus and the "Balanced Annotated Collection of Texts" (SAT-corpus). The research involves methodologies of different linguistics branches (phonetics, discourse analysis, sociolinguistics, psycholinguistics, corpus linguistics, etc.), making it possible to built formal statistical schemes which may be used both for theoretical linguistic studies and the improvement of NLP tasks.

MOTS-CLÉS: discours parlé, recherche interdisciplinaire, pragmatique, marqueurs pragmatiques, le russe, linguistique de corpus, sociolinguistique

KEYWORDS: spoken discourse, interdisciplinary research, pragmatics, pragmatic markers, Russian, corpus linguistics, sociolinguistics

1 Pragmatic markers

Pragmatic markers (PMs) are discourse units (words and multiword expressions) with a weakened referential meaning, which perform a variety of pragmatic tasks. These speech elements form a mandatory component of oral communication in any language (Ghezzi & Molinelli 2014) – they allow to break the discourse flow into fragments, help the speaker to establish correct relationship with the interlocutor, help to convey the speaker's stance and perform many other pragmatic functions. Therefore, it is largely PMs that are responsible for the effectiveness of communication. However, unlike other lexical units that are well-represented in numerous dictionaries, pragmatic markers for many languages are very poorly studied.

It can be assumed that PMs may be observed in every natural spoken language. For example, in English the common PMs are "well", "you know", "I don't know", and many others, and in Russian the similar PMs are "tak" ("well"), "znaesh" ("you know"), "ne znaju" ("I don't know"), etc. In earlier papers on spoken discourse, PM were considered within a wider class of *discourse particles* or *discourse markers*.

The term "pragmatic marker" was first introduced by B. Fraser, who proposed an extensive classification of units defined as "the linguistically encoded clues which signal the speaker's potential communicative intentions" (Fraser 1996). According to Fraser, such a functional class of expressions exists in any language. Pragmatic markers are embedded in the sentence, but are "separate and distinct from the propositional content of the sentence". They signal important aspects of the speaker's message (Fraser 1990).

In this paper we adhere to understanding PM as it was proposed in (Bogdanova-Beglarian 2014), and the distinction between pragmatic markers (PMs) and discourse markers (DMs) as we see it is presented in (Bogdanova-Beglarian et al. 2018).

Pragmatic markers form a considerable part of the general "verbal mass" of oral discourse, therefore understanding the patterns of their functioning is necessary for any research whose object is unprepared spoken speech. The need to PM study is determined by the importance of studying natural oral speech for linguistic research (primarily, for phonetic and grammatical studies of spontaneous speech, as well as for studying general patterns of spoken discourse) and, more broadly, for researching people's speech behavior in the society.

2 Interdisciplinary approach to PMs study

The current research demonstrates an interdisciplinary approach to study of PMs based on two representative speech corpora – ORD corpus of Russian Everyday Speech known as "One-Day-of-Speech"-corpus and the "Balanced Annotated Collection of Texts" (SAT-corpus).

The first one is the corpus of Russian Everyday Speech "One Day of Speech" (ORD corpus), which is today one of the most representative resources for analysis of Russian oral discourse (Asinovsky et al. 2009). An important feature of this resource is the fact that similar to the principle of "Holter monitoring", which is widely used in cardiology to record heart's rate and rhythm, volunteer-informants recorded their entire speech communication during one day. Thereby, sound recordings of everyday speech in a natural situation were obtained (ibid.). This corpus was recently subjected to sociological extension, in the process of which we have gathered recordings from new participants, belonging to different gender, age, status and professional groups (service workers, white-collars, IT-specialists, university professors, artists, musicians, businessmen, pensioners, students, etc.) (Bogdanova-Beglarian et al. 2017). Now, the ORD collection exceeds 1400 h of recordings, presenting speech of 128 respondents and a thousand of their interlocutors.

The second corpus is the so-called "Balanced Annotated Collection of Texts" (SAT) which includes monologue speech recordings received from different professional groups of native speakers. All texts in SAT were obtained in 4 experiments – reading, retelling, image description, storytelling (Bogdanova-Beglarian et al. 2017). Participants were balanced in terms of their social characteristics. This speech collection contains several modules of speech obtained from homogeneous professional groups: 1) doctors' speech (MED); 2) speech of lawyers (JUR); 3) speech of Russian teachers (RKI), 4) speech of students (STUD), 5) speech of IT-specialists (COMP), etc.

Pragmatic markers annotation is performed in the ELAN multimedia annotation software. It consists of the following elements: pragmatic markers allocation, PM type and functions, morphological and grammatical markup, and the characteristics of PM phonetic features. To obtain statistical information on pragmatic markers distribution we use methods of quantitative linguistics, in particular, statistical methods for received annotations processing (descriptive statistics, multivariate analysis, cluster analysis, etc.) and testing of statistical hypotheses.

The research involves methodologies of different linguistics branches (e. g., phonetics, discourse analysis, sociolinguistics, psycholinguistics, corpus linguistics, etc.). For example, sociolinguistic analysis is performed based on the results of the questionnaire, which was taken by all participants in the sound recording. Psycholinguistic analysis is based on the results of standard psychological tests, which were performed by informants at the day of recording. Pragmatic studies may be enriched by information about the conditions and the type of the communicative scenario imported from the ORD corpus.

3 Some results and applications

Processing the results of the pilot annotation allowed to obtain preliminary data on frequency of individual pragmatic markers and their types, as well as on the dependence of PM usage on gender, age, psychological type, temperament, and the level of speech competence of the speaker. As a result of statistical data processing, frequency lists of both PMs and their functions were obtained. Thus, according to the results obtained from the ORD corpus of everyday Russian (10,11), their share can reach up to 6% of the total number of words in speech of individual speakers. More than that, in some speech fragments, PMs may even exceed the share of significant units (i. e., standard words).

Currently, we proceed continuous annotation for the both corpora. It is expected that the obtained results will be of great importance primarily for the phonetic, lexical, syntactic and pragmatic studies of oral discourse, the study of collocations and idioms of spoken speech. In addition, the results have value for sociolinguistics, cognitive linguistics, discourse analysis, practical rhetoric, linguistic anthropology, Russian language pedagogy, and for other scientific

disciplines related with the examination of oral discourse and the study of its patterns, as well as for the study of speech behavior as one of the basic forms of people social behavior.

The results of the project will find their practical application: 1) in the field of the applied linguistics, informational and speech technologies – to support the systems of automatic speech monitoring, voice search, speech synthesis and recognition systems, artificial intelligence, voice dialog systems when communicating with a computer or robot, 2) for teaching Russian as a foreign language, 3) for conducting linguistic and forensic expertise based on audio records of speech communication.

Acknowledgements

The presented research is supported by the Russian Science Foundation, project #18-18-00242 "Pragmatic Markers in Russian Everyday Speech".

References

Asinovsky, A., Bogdanova, N., Rusakova, M., Ryko, A., Stepanova, S., and Sherstinova, T. (2009) The ORD Speech Corpus of Russian Everyday Communication "One Speaker's Day": Creation Principles and Annotation. In: Matoušek, V., Mautner, P. (eds.) *TSD 2009. LNCS*, vol. 5729, pages 250–257. Springer, Heidelberg.

Ghezzi, C., Molinelli, P. (eds) (2014) *Discourse and Pragmatic Markers from Latin to the Romance Languages*. Oxford: Oxford University Press.

Fraser, B. (1990) An approach to discourse markers, *Journal of Pragmatics*, vol. 14, 1990, pages 383–395.

Fraser, B. (1996) Pragmatic markers. In *Pragmatics*, vol. 6, is. 2, pages 167–190.

Bogdanova-Beglarian, N. V. (2014). Pragmatemy v ustnoj povsednevnoj rechi: opredelenie ponyatia i obshchaja tipologia ["Pragmatic markers in spoken everyday speech: Definition and general typology"]. In *Vestnik Permskogo universiteta. Rossijskaja i zarubezhnaja filologia* [*Perm University Herald. Russian and Foreign Philology*], iss. 3 (27), pages 7–20.

Bogdanova-Beglarian, N., Blinova, O., Martynenko, G., Sherstinova, T., Zaides, K. (2018): Pragmatic Markers in Russian Spoken Speech: an Experience of Systematization and Annotation for the Improvement of NLP Tasks. In *Proc. of the FRUCT'23*. Bologna, Italy, 13–16 Nov. 2018 / S. Balandin et al. (eds.). FRUCT Oy. Finland, pages 69–77.

Bogdanova-Beglarian, N., Sherstinova, T., Zajdes, K. (2017): Corpus "Balanced Annotated Text Library": Methodology Multi-Level Analysis of the Russian Monological Speech [Korpus "Sbalansirovannaja Annotirovannaja Tekstoteka": metodika mnogourovnevogo analiza russkoj monologicheskoj rechi]. In *Analysis of Spoken Russian (AR3-2017). Proc. of the 17th Int. Seminar*, St. Petersburg, pages 8–13.

Bogdanova-Beglarian N., Sherstinova T., Blinova O., Ermolova O., Baeva E., Martynenko G., and Ryko A. (2016) Sociolinguistic Extension of the ORD Corpus of Russian Everyday Speech. In *SPECOM 2016, Lecture Notes in Artificial Intelligence, LNAI*, vol. 9811. Springer, Switzerland, pages 659–666.

Daba software for written corpora of underresourced languages

Kirill Maslinsky[1], Valentin Vydrin[2, 3]
(1) Higher School of Economics, St. Petersburg, Russia
(2) Inalco, Paris, France
(3) LLACAN (CNRS, UMR-8135), Paris, France
maslinych@gmail.com, jjmeric@gmail.com, vydrine@gmail.com

Résumé

Le logiciel Daba a été développé pour la création des corpus écrits pour des langues peu dotées. Il a été appliqué aux corpus de plusieurs langues de la famille mandé. Les outils du Daba sont : le Metaeditor (une interface pour introduction des metadonnées des textes), l'analyseur morphologique Gpars (le Daba utilise le modèle d'annotation inspiré par le glosage des exemples dans les publications linguistiques), l'interface de désambiguïsation Gdisamb. Les défis de la création des corpus des langues peu dotées sont discutés : le manque de standardisation des versions écrites des langues ; la rareté des textes disponibles sous formats électroniques; la nécessité de développer les interfaces conviviaux en minimisant l'assistance d'un informaticien.

Abstract

The Daba software has been developed for corpus building for under-resourced languages and used for written corpora of several languages of the Mande family. The Daba tools are represented: the Metaeditor (an interface for adding metadata to the texts), the morphological analyser Gpars (Daba uses a morpheme-based morphological annotation scheme inspired by interlinear glossed form of presentation of linguistic examples), the disambiguation interfacte Gdisamb. Specific challenges for the corpus-building for underresourced languages are discussed: a low level of written language standardization; limited amount of available texts in the electronic format; a necessity to develop a user-friendly software which would need only limited assistance by a programmer.

Mots-clés : langue peu dotées, analyseur morphologique, désambiguïsation, corpus écrit.

Keywords: under-resourced language, morphological analyzer, disambiguation, corpus-building, written corpus

1 Written corpora for underresourced languages

Corpus building for a language with low level of literacy and language standardization is different from the corpora of languages with long-established standards for orthography, lexical choice, and linguistic description. The following peculiarities can be mentioned:

– limited amount of texts available, especially in electronic formats. As a rule, such languages a underrepresented in the Internet;

- low level of standardization, sometimes competing orthographic systems;
- often insufficient grammatical and lexicographical description;
- necessity of glosses.

2 Corpus building procedures. Daba software package and other tools used in the Corpora Mandeica project

Daba has been developed by Kirill Maslinsky originally for the Bambara Reference Corpus (since 2009), then adapted to some other languages of Mande family (Maninka, Eastern Dan, Mwan; there are ongoing projects for the corpora of Guinean Kpelle and Mano). The online search interface used for these corpora is NoSketchEngine (which is to be replaced by the Context interface).

The corpus building procedure includes the following stages.

1. Digitalizing texts available only on paper; collecting texts available in the internet.
2. Standardization of orthography; development of convertors from non-standard to the standard orthographies (to the extent of possible).
3. Development a lexical database (or adapting an existing dictionary) in the Toolbox format, in conformity with certain conventions.
4. Adding metadata to the texts (the Metaeditor interface).
5. Automated morphological annotation of all texts (the gpars interface). The forms are identified (i.e., their correspondences are found in the lexical database) and lemmatized; they are attributed with POS markers and glosses. For compound words, a multylayer glossing is provided.
6. Manual disambiguation of selected texts (the gdisamb interface).
7. Phrase-by-phrase synchronization of the texts (for parallel corpora; the gdisamb interface).
8. Syntactic tagging (UD-format).
9. Creating online search interface with flexible possibilities for concordance building (NoSketchEngine, in perspective: Context).

The Daba package is written in the Python format. It includes the Metaeditor, a morphological analyzer (gparser), and the disambiguation interface (gdisamb).

The outcomes: small- and medium-size relatively low-cost corpora for underresourced languages. Parallel corpora. In perspective: oral corpora.

3 Further challenges

The main challenge is a further development of the Daba package in order to make it user-friendlier, in order to gradually minimize the assistance on the part of the programmer (so far, building of a new corpus is impossible without such an assistance).

Références

Maslinsky, Kirill (2014): Daba: a model and tools for Manding corpora. In: *21ème Traitement Automatique des Langues Naturelles*, pp. 114–122.

Rovenchak, Andrij (2018). Texts for the corpus of Nko: collection, conversion, and open issues. *Mandenkan* 59, p. 57-66.

Vydrin, Valentin (2013): Bamana Reference Corpus (BRC). In: *Procedia Social and Behavioral Sciences* (95), S. 75–80.

Valentin Vydrin (2014). Projet des corpus écrits des langues manding : le bambara, le maninka. In : Mathieu Mangeot, Fatiha Sadat (éd.). Actes de l'atelier sur le traitement automatique des langues africaines TALAf 2014. (Actes des Ateliers TALN 2014. Éd. par Brigitte Bigi, ISBN : 978-2-9518233-6-5) http://www.taln2014.org/site/actes-en-ligne/actes-en-ligne-ateliers/

Vydrin, Valentin (2018). Corpus-driven lexicography for African languages: Perspectives for Manding. In: *The 9th World Congress Of African Linguistics: African languages in a global world: from description to state policies*. Rabat, Mohammad V University of Rabat, 25-28.08.2018, p. 75.

Vydrin, Valentin & Rovenchak, Andrij & Maslinsky, Kirill (2016). Maninka Reference Corpus: A Presentation. In: TALAf 2016 : Traitement automatique des langues africaines (écrit et parole) Atelier JEP-TALN-RECITAL 2016 - Paris le 4 juillet 2016. Actes. 8 p. http://talaf.imag.fr/2016/Actes/VYDRIN_ET_AL%20-%20Maninka%20Reference%20Corpus: %20A%20Presentation.pdf

Combiner parseur automatique et révision manuelle pour la constitution d'un corpus arboré de parole spontanée : retour d'expérience sur le corpus ODIL_syntaxe

Ilaine Wang[1, 2] Aurore Pelletier[2] Jakub Waszcuk[3] Anaïs Halftermeyer[1]
Jean-Yves Antoine[2] Lotfi Abouda[4] Emmanuel Schang[4] Agata Savary[2]
(1) Laboratoire LIFO, U. Orléans, 45000 Orléans, France
(2) Laboratoire LIFAT, U. Tours, 37000 Blois, France
(3) ISI, Heinrich-Heine-Universität Düsseldorf, 40225 Deutschland
(4) LLL, CNRS U. Orléans, 45000 Orléans, France
ilaine.wang,anais.halftermeyer@univ-orleans.fr,
jakub.waszczuk@phil.uni-duesseldorf.de,jean-yves.antoine@univ-tours.fr

RÉSUMÉ

Cet article présente l'utilisation d'une plateforme d'annotation syntaxique (*Contemplata*) qui intègre un parseur pour annoter automatiquement des corpus écrits ou oraux puis permettre leur révision manuelle par un·e expert·e, afin de limiter son travail d'annotation. Dans le cadre du projet ODIL, cet outil a permis de réaliser un corpus de français parlé spontané annoté en arbres de constituants, ceci dans la perspective d'une annotation en temporalité. Nous présentons ici la démarche mise en œuvre pour l'annotation ainsi que les conventions, et proposerons une démonstration de l'outil.

ABSTRACT

Combining Automatic Parsing and Manual Revision for the Constitution of a Spontaneous Speech Treebank : Experience Feedback on the ODIL_Syntaxe Corpus.

This paper describes a syntactic annotation platform (*Contemplata*) that integrates a parser (*Stanford Parser* precisely) to automatically annotate written text or oral transcriptions and then allows their manual revision by an expert, in order to ease the annotation task. This tool was used in the ODIL Project to produce a phrase-structure treebank based on a corpus of spontaneous speech. In this paper, we present the annotation process that has been implemented as well as our annotation guidelines and plan to provide a demonstration of the annotation tool during the presentation.

MOTS-CLÉS : annotation syntaxique semi-automatique, corpus arboré, arbres de constituants, outil d'annotation, parseur, français parlé spontané.

KEYWORDS: syntactic annotation, treebank, phrase-structure representation, annotation tool, parser, spontaneous spoken French.

1 Projet ODIL : annotation en temporalité

Le travail qui sera présenté avec ce poster a été réalisé dans le cadre du lot « annotation temporelle » du projet ODIL, financé par la région Centre Val de Loire. Le sous-projet Temporal@ODIL vise la réalisation d'un corpus oral annoté en temporalité (identification des éventualités et caractérisation

des relations temporelles entre éventualités) qui n'a d'équivalent pour le français qu'une seule ressource : le *French TimeBank* (Bittar et al., 2011). L'originalité de nos travaux est précisément de se focaliser, contrairement à ce dernier, sur l'oral spontané. Une première partie du projet a consisté à étudier l'adaptation au langage oral de la norme d'annotation de la temporalité ISO TimeML (Pustejovsky et al., 2003). Cette extension (Antoine et al., 2017) de la norme consiste en particulier à décrire la temporalité non plus au niveau de la tête lexicale des éventualités, mais au niveau des nœuds d'une représentation arborée des énoncés. La considération de la structure syntaxique des énoncés autorise en effet une représentation parfois nécessaire de l'empan des éventualités, tout en facilitant la tâche de l'annotateur. Par ailleurs, une ambiguité syntaxique peut nuire à l'annotation sémantique, une fois séparés les niveaux d'annotation, la charge cognitive de l'annotateur s'en trouve considérablement réduite. Les éventualités correspondent en effet toujours à un nœud de l'arbre syntaxique, dès lors que celui-ci relève d'une représentation en constituants et non en dépendances.

Ce choix nous a donc conduit à élaborer au préalable un corpus arboré en constituants. Cet article présente précisément notre démarche d'annotation syntaxique. Un des intérêts de notre démarche est de conduire une annotation semi-automatique : nous utilisons en effet un analyseur syntaxique reconnu, le *Stanford Parser* (Green et al., 2011), pour proposer une première annotation qui sera ensuite révisée manuellement. Pour profiter au mieux de cette stratégie, nous avons apporté un soin très particulier à l'utilisabilité du nouvel outil d'annotation que nous avons développé pour le projet : *Contemplata*. Il s'agit d'un outil générique qui permet d'annoter tout corpus en arbres de constituants mais aussi de décorer les nœuds de cet arbre et d'y ajouter des relations de natures variées. Cet article présente un retour d'expérience sur l'utilisation de l'outil dans le cadre d'ODIL.

2 Corpus ODIL_Syntaxe

Le projet ODIL fournira, à sa clôture, un corpus de français parlé de 12355 mots qui comportera (1) une annotation en arbres de constituants et (2) une annotation en relations temporelles. Le corpus arboré, *ODIL_Syntaxe*, est désormais finalisé et sera distribué sous licence libre d'ici à fin 2019. Il repose sur l'enrichissement de trois corpus bruts de transcriptions de l'oral correspondant à trois registres différents : les corpus ESLO (entretiens sociolinguistiques issus de Eshkol Taravella et al. (2011)), OTG (dialogue oral en présentiel) et Accueil_UBS (dialogue oral par téléphone). Ces deux dernières ressources ont déjà servi à la réalisation d'ANCOR_Centre, un très grand corpus de français parlé annoté en coréférence (Muzerelle et al., 2014).

La conception de corpus annotés en arbres syntaxiques a pris un nouvel essor depuis le début du millénaire. De telles ressources sont cruciales pour l'apprentissage de systèmes de traitement automatique des langues dans des applications nécessitant la prise en compte de la structure syntaxique des énoncés. Elles constituent également un préalable pour les études linguistiques en corpus opérant à ce niveau de description linguistique (coréférence, saillance discursive, relations temporelles. . .). Le français dispose déjà de ressources importantes avec le *French Treebank* (Abeillé et al., 2003) et *SEQUOIA* (Candito and Seddah, 2012) mais leur confection s'est faite à partir de textes écrits et les résultats ne sont pas transposables sur les réalisations orales. *SEQUOIA* propose par ailleurs une annotation en dépendances et non pas en constituants. A un moment où l'ingénierie des langues et la linguistique de corpus sont confrontées de manière croissante à des contenus de parole,

ce corpus arboré se propose de remédier à cette lacune. Un seul corpus peut être comparé au nôtre : *Rhapsodie* (Lacheret et al., 2014), dont l'accès pour la version en constituance est toutefois limité.

Les conventions d'annotation du corpus reposent principalement sur celles du *French TreeBank* telles qu'utilisées pour l'apprentissage du *Stanford Parser*. Nous avons toutefois dû ajouter certaines règles pour les adapter à l'oral spontané. Ces adaptations ont essentiellement concerné la représentation des disfluences orales, en particulier les inachèvements et toutes les formes d'entassement paradigmatique (reprises, répétitions, autocorrections). Pour ces adaptations, nous avons cherché à rester aussi proches que possible des choix d'annotation de Rhapsodie, qui nous sont apparues des plus pertinentes.

3 Contemplata : outil d'annotation syntaxique semi-manuelle

Contemplata se présente sous la forme d'une application Web dont la partie client est développée en langage Elm puis recompilé en JavaScript et utilisable dans tout navigateur Web sans aucune installation spécifique. Le client communique avec un serveur développé en Haskell, qui permet en particulier l'interrogation du *Stanford Parser* ou tout autre outil équivalent respectant les formats d'entrée et de sortie des données manipulées. Au final, l'application Web permet de gérer l'intégralité de la réalisation d'un corpus arboré, en passant de la supervision du projet d'annotation (rôle administrateur) à l'annotation en elle-même (rôle d'annotateur ou d'adjudicateur). En pratique, l'annotation suit une succession d'étapes que l'annotateur doit respecter, avant tout pour lui simplifier la tâche. Cette décomposition en sous-tâches réduit en effet la charge cognitive de l'expert·e, et favorise la qualité de l'annotation. L'annotation respecte la succession suivante :

— Analyse syntaxique automatique par le *Stanford Parser*, auquel nous avons ajouté une étape de pré-traitement qui met à l'écart (sans les éliminer) les formules d'introduction et les phatiques, qui sont très présents en oral spontané tout en ne portant pas de contenu propositionnel.
— Correction manuelle de l'annotation automatique en parties du discours (POS) et élimination des phatiques résiduels encore présents.
— Relance de l'analyse syntaxique automatique après correction des POS. Cette première correction permet en effet souvent au parseur de corriger l'ensemble de son analyse.
— Segmentation manuelle des tours de paroles en plusieurs énoncés. Les longs tours de parole peuvent induire en erreur le parseur. C'est notamment le cas des registres de dialogue du type « entretien » (ESLO). On découpe donc les tours de parole en énoncés successifs cohérents.
— Relance de l'analyse syntaxique automatique, le découpage en énoncé pouvant, une fois encore, améliorer la qualité des arbres obtenus automatiquement.
— Correction manuelle des arbres résultants. Il s'agit de l'opération la plus coûteuse humainement, mais les étapes précédentes ont amélioré la qualité de l'annotation automatique, ce qui réduit la charge de l'expert·e. *Contemplata* permet toutes les modifications possibles des arbres syntaxiques : ajout, déplacement et suppression de nœuds, déplacement de sous-arbres, modification de l'étiquette d'un nœud etc. (Figure 1). Cette étape gère également l'annotation manuelle des disfluences orales, qui ne sont pas prises en compte par le parseur, puisque celui-ci a été entraîné sur un corpus d'écrit.

La réalisation du corpus *ODIL_Syntaxe* a montré que l'annotation est grandement facilitée par *Contemplata*, ce qui nous a permis d'annoter un corpus de taille déjà raisonnable (comparable au *French TimeBank*) à un coût humain relativement limité. Notons enfin que *Contemplata* permet une

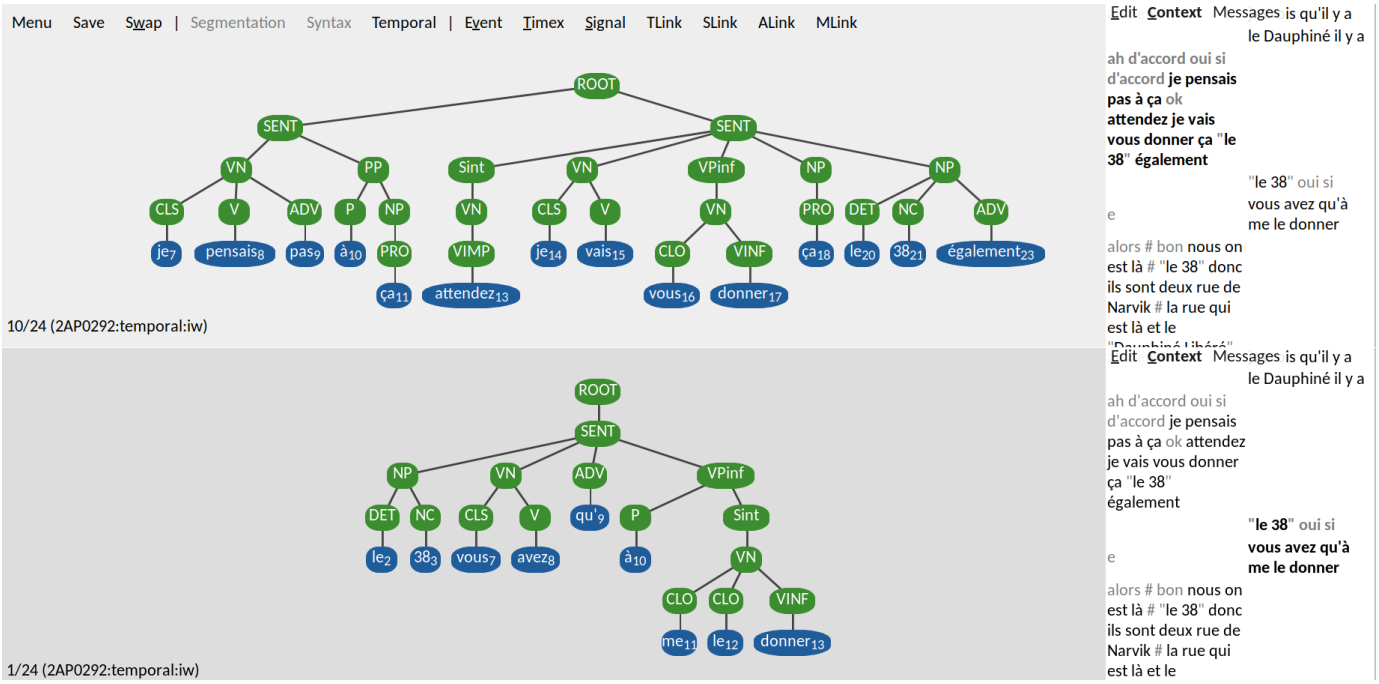


FIGURE 1 – Interface de *Contemplata* affichant en simultané la structure syntaxique en constituant de deux tours de parole (ROOT) comportant ici un ou deux énoncés (SENT)

révision experte de l'annotation obtenue par le biais d'une aide à l'adjudication d'annotations (comparaison graphique de 2 annotations concurrentes sur un même extrait de corpus).

Références

Abeillé, A., Clément, L., and Toussenel, F. (2003). Building a treebank for French. In *Treebanks*.

Antoine, J.-Y., Waszczuk, J., Lefeuvre Halftermeyer, A., Abouda, L., Schang, E., and Savary, A. (2017). Temporal@ ODIL Project : Adapting ISO-TimeML to Syntactic Treebanks for the Temporal Annotation of Spoken Speech. In *Proceedings isa-13 (IWCS'2017)*.

Bittar, A., Amsili, P., and Denis, P. (2011). French TimeBank : un corpus de référence sur la temporalité en français. In *Actes de TALN'2011*, pages 259–270.

Candito, M. and Seddah, D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Actes de TALN'2011*, pages 321–334.

Eshkol Taravella, I., Baude, O., Maurel, D., Hriba, L., Dugua, C., and Tellier, I. (2011). Un grand corpus oral « disponible » : le corpus d'Orléans 1 1968-2012. *TAL*, 52(3) :17–46.

Green, S., De Marneffe, M.-C., Bauer, J., and Manning, C. D. (2011). Multiword expression identification with tree substitution grammars : A parsing tour de force with french. In *Proceedings EMNLP'2011*.

Lacheret, A., Kahane, S., Beliao, J., Dister, A., Gerdes, K., Goldman, J.-P., Obin, N., Pietrandrea, P., and Tchobanov, A. (2014). Rhapsodie : un Treebank annoté pour l'étude de l'interface syntaxe-prosodie en français parlé. In *Actes de CMLF'2014*, pages 2675–2689.

Muzerelle, J., Lefeuvre, A., Schang, E., Antoine, J.-Y., Pelletier, A., Maurel, D., Eshkol, I., and Villaneau, J. (2014). ANCOR_Centre, a large free spoken French coreference corpus : description of the resource and reliability measures. In *Proceedings LREC'2014*.

Pustejovsky, J., Castano, J. M., Ingria, R., Sauri, R., Gaizauskas, R. J., Setzer, A., Katz, G., and Radev, D. R. (2003). TimeML : Robust specification of event and temporal expressions in text. *New directions in question answering*, 3 :28–34.

How Does Language Influence Documentation Workflow? Unsupervised Word Discovery Using Translations in Multiple Languages

Marcely Zanon Boito[1] Aline Villavicencio[2,3] Laurent Besacier[1]
(1) Laboratoire d'Informatique de Grenoble (LIG), UGA, G-INP, CNRS, INRIA, France
(2) Department of Computer Science, University of Sheffield, England
(3) Institute of Informatics (INF), UFRGS, Brazil
contact: marcely.zanon-boito@univ-grenoble-alpes.fr

RÉSUMÉ

Comment la langue influence le processus de documentation ? Découverte non supervisée de mots basée sur des traductions en langues multiples

Pour la documentation des langues, la transcription est un processus très coûteux : une minute d'enregistrement nécessiterait environ une heure et demie de travail pour un linguiste (Austin and Sallabank, 2013). Récemment, la collecte de traductions (dans des langues bien documentées) alignées aux enregistrements est devenue une solution populaire pour garantir l'interprétabilité des enregistrements (Adda et al., 2016) et aider à leur traitement automatique. Dans cet article, nous étudions l'impact de la langue de traduction sur les approches automatiques en documentation des langues. Nous traduisons un corpus parallèle bilingue Mboshi-Français (Godard et al., 2017) dans quatre autres langues, et évaluons l'impact de la langue de traduction sur une tâche de segmentation en mots non supervisée. Nos résultats suggèrent que la langue de traduction peut influencer légèrement la qualité de segmentation. De plus, combiner l'information apprise par différents modèles bilingues nous permet d'améliorer ces résultats.

ABSTRACT

For language documentation initiatives, transcription is an expensive resource: one minute of audio is estimated to take one hour and a half on average of a linguist's work (Austin and Sallabank, 2013). Recently, collecting aligned translations in well-resourced languages became a popular solution for ensuring posterior interpretability of the recordings (Adda et al., 2016). In this paper we investigate language-related impact in automatic approaches for computational language documentation. We translate the bilingual Mboshi-French parallel corpus (Godard et al., 2017) into four other languages, and we perform bilingual-rooted unsupervised word discovery. Our results hint towards an impact of the well-resourced language in the quality of the output. Moreover, by combining the information learned by different bilingual models, we are able to increase the quality of the segmentation.

MOTS-CLÉS : découverte non supervisée du lexique, documentation des langues, approches multilingues.

KEYWORDS: unsupervised word discovery, language documentation, multilingual approaches.

1 Introduction

The *Cambridge Handbook of Endangered Languages* (Austin and Sallabank, 2011) estimates that at least half of the 7,000 languages currently spoken worldwide will no longer exist by the end of this century. For these *endangered* languages, data collection campaigns have to accommodate the challenge that many of them are from oral tradition, and producing transcriptions is costly. This *transcription bottleneck* problem can be handled by translating into a widely spoken language to ensure subsequent interpretability of the collected recordings, and such parallel corpora have been recently created by aligning the collected audio with translations in a well-resourced language (Adda et al., 2016; Godard et al., 2017; Boito et al., 2018). Moreover, some linguists suggested that more than one translation should be collected to capture deeper layers of meaning (Evans and Sasse, 2004).

This work is a contribution to the Computational Language Documentation (CLD) research field, that aims to replace part of the manual steps performed by linguists during language documentation initiatives by automatic approaches. Here we investigate the unsupervised word discovery and segmentation task, using the bilingual-rooted approach from Godard et al. (2018). There, words in the well-resourced language are aligned to unsegmented phonemes in the endangered language in order to identify group of phonemes, and to cluster them into word-like units. We experiment with the Mboshi-French parallel corpus, translating the French text into four other well-resourced languages in order to investigate language impact in this CLD approach. Our results confirm the extensive investigation performed in Boito et al. (2019b), showing that this language impact exists, and that models based on different languages will output different word-like units.

2 Methodology

The Multilingual Mboshi Parallel Corpus: In this work we extend the bilingual Mboshi-French parallel corpus (Godard et al., 2017), fruit of the documentation process of Mboshi (Bantu C25), an endangered language spoken in Congo-Brazzaville. The corpus contains 5,130 utterances, for which it provides audio, transcriptions and translations in French. We translate the French into four other well-resourced languages through the use of the $DeepL$ translator.[1] The languages added to the dataset are: English, German, Portuguese and Spanish. Table 1 shows some statistics for the produced *Multilingual Mboshi* parallel corpus.[2]

Bilingual Unsupervised Word Segmentation/Discovery Approach: We use the bilingual neural-based Unsupervised Word Segmentation (UWS) approach from Godard et al. (2018) to discover words in Mboshi. In this approach, Neural Machine Translation (NMT) models are trained between language pairs, using as source language the translation (word-level) and as target, the language to document (unsegmented phonemic sequence). Due to the attention mechanism present in these networks (Bahdanau et al., 2014), posterior to training, it is possible to retrieve *soft-alignment probability matrices* between source and target sequences. These matrices give us sentence-level source-to-target alignment information, and by using it for clustering neighbor phonemes aligned to the same translation word, we are able to create segmentation in the target side. The product of this approach is a set of (discovered-units, translation words) pairs.

[1] Available at `https://www.deepl.com/translator`

[2] Available at `https://github.com/mzboito/mmboshi`

	MB	FR	EN	ES	DE	PT
# Types	6,633	5,178	4,392	5,473	5,641	5,465
# Tokens	30,556	42,715	37,379	37,428	37,515	37,095
Avg. Token Length	4.18	4.41	4.19	4.36	4.91	4.40
Avg. Tokens/Sentence	5.96	8.33	7.29	7.30	7.31	7.23

Table 1: Statistics for the Multilingual Mboshi parallel corpus. The French text is used for generating translation in the four other languages present in the right side of the table.

Bilingual			Multilingual Voting					ANE Selection
1	FR	73.40		25%	50%	75%	100%	
2	EN	73.10	(1-2)	73.10	73.10	73.30	73.30	73.80
3	PT	72.80	(1-3)	72.40	74.60	72.10	72.10	73.90
4	ES	72.60	(1-4)	71.60	74.80	74.20	70.90	73.90
5	DE	71.00	(1-5)	74.30	74.90	73.10	70.00	73.90

Table 2: From left to right, results for: bilingual UWS, multilingual leveraging by voting, ANE selection.

Multilingual Leveraging: In this work we apply two simple methods for including multilingual information into the bilingual models from Godard et al. (2018). The first one, **Multilingual Voting**, consists of merging the information learned by models trained with different language pairs by performing a voting over the final discovered boundaries. The voting is performed by applying an agreement threshold T over the output boundaries. This threshold balances between accepting all boundaries from all the bilingual models (zero agreement) and accepting only input boundaries discovered by all these models (total agreement). The second method is **ANE Selection**. For every language pair and aligned sentence in the dataset, a soft-alignment probability matrix is generated. We use *Average Normalized Entropy* (ANE) (Boito et al., 2019a) computed over these matrices for selecting *the most confident one* for segmenting each phoneme sequence. This exploits the idea that models trained on different language pairs will have language-related behavior, thus differing on the resulting alignment and segmentation over the same phoneme sequence.

3 Experiments

The experiment settings from this paper follow Boito et al. (2019b), while the evaluation protocol for the Mboshi corpus (Boundary F-scores using the ZRC speech reference) is the same from Boito et al. (2019a). Table 2 presents the results for bilingual UWS and multilingual leveraging. For the former, we reach our best result by using as aligned information the French, the original aligned language for this dataset. Languages closely related to French (Spanish and Portuguese) ranked better, while our worst result used German. English also performs notably well in our experiments, what was also observed in previous work (Boito et al., 2019b). We believe this is due to the statistics features of the resulting text. We observe in Table 1 that the English portion of the dataset contains the smallest vocabulary among all languages. Since we train our systems in very low-resource settings, vocabulary-related features can impact greatly the system's capacity to language-model, and consequently the final quality of the produced alignments. Even in high-resource settings, it was already attested that some languages are more difficult to model than others (Cotterell et al., 2018).

For the multilingual selection experiments, we experimented combining the languages from top to

MB-DE		MB-EN		MB-ES		MB-FR		MB-PT	
itua	itoua	ibara	ibara	ingobha	ingobha	itua	itoua	oboá+ngá	oboa
mɯndzɯ	monzo	otsɛ́ngɛ	otsenge	ondóngo	ondongo	itúa+ngá	itoua	**ERROR**	nyaamvua
tsimba	tsimba	asúa	asoua	mbía+mbvúlá	amvoulou	itúa+mbía	itoua	itua	itoua
abia	Freunde	okúmú	okoumou	itua	itoua	kánga	pintade	mbembe	mbembe
tsósá	Henne	olangi	bottle	y'+konga	cuerno	**ERROR**	nyobhosi	tsimba	tsimba
ibara	Ibara	tsɛ́si	hare	itúa+ngá	itoua	oboá+ngá	oboa	okwɯɯ	cordão
andzɯi	Elefanten	itua	itoua	oboá+ngá	oboa	kyéma	singe	mómeá	tentar
ondúma	Onduma	kóli	badger	ekoko	ekoko	**ERROR**	amassez	abvúɛ	cunhado
ikinyi	Fliege	andzúe	bees	okubha	herrero	tsimba	tsimba	ekoko	ekoko
itúa+ngá	itoua	itúa+mbía	itoua	ibara	ibara	lekú+yá	guêpes	mbósi	cabras

Table 3: Top 10 confident (discovered type, translation) pairs for the five bilingual models. The "**+**" mark means the discovered type is a concatenation of two existing true types.

bottom as they appear Table 2 (ranked by performance; e.g. 1-3 means the combination of FR(1), EN(2) and PT(3)). We observe that the performance improvement is smaller than the one observed in Boito et al. (2019b), what we attribute to the fact that our dataset was artificially augmented. This could result in the available multilingual form of supervision not being as rich as in a manually generated dataset. Finally, the best boundary segmentation result is obtained by performing multilingual voting with all the languages and an agreement of 50%, what indicates that the information learned by different languages will provide additional complementary evidence.

Lastly, following the methodology from Boito et al. (2019a), we extract the most confident alignments (in terms of ANE) discovered by the bilingual models. Table 3 presents the top 10 most confident (discovered type, translation) pairs. The Mboshi phoneme sequences were replaced by their grapheme equivalents to increase readability, but all results were computed using phonemes. Looking at the pairs the bilingual models are most confident about, we observe there are some types discovered by all the bilingual models (e.g. Mboshi word *itua*, and the concatenation *oboá+ngá*). However, the models still differ for most of their alignments in the table. This hints that while a portion of the lexicon might be captured independently of the language used, other structures might be more dependent of the chosen language. On this note, Haspelmath (2011) suggests the notion of *word* cannot always be meaningfully defined cross-linguistically.

4 Conclusion

In this work we follow the investigation from Boito et al. (2019b), training bilingual UWS models using the endangered language Mboshi as target and different well-resourced languages as aligned information. Results show that similar languages rank better in terms of segmentation performance, and that by combining the information learned by different models, segmentation is further improved. This might be due to the different language-dependent structures that are captured by using more than one language. Lastly, we extend the bilingual Mboshi-French parallel corpus, creating a multilingual corpus for the endangered language Mboshi that we make available to the community.

References

Adda, G., Stüker, S., Adda-Decker, M., Ambouroue, O., Besacier, L., Blachon, D., Bonneau-Maynard, H., Godard, P., Hamlaoui, F., Idiatov, D., Kouarata, G.-N., Lamel, L., Makasso, E.-M., Rialland, A., de Velde, M. V., Yvon, F., and Zerbian, S. (2016). Breaking the unwritten language barrier: The BULB project. *Procedia Computer Science*, 81:8–14.

Austin, P. K. and Sallabank, J. (2011). *The Cambridge handbook of endangered languages*. Cambridge University Press.

Austin, P. K. and Sallabank, J. (2013). *Endangered languages*. Taylor & Francis.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Boito, M. Z., Anastasopoulos, A., Lekakou, M., Villavicencio, A., and Besacier, L. (2018). A small griko-italian speech translation corpus. *arXiv preprint arXiv:1807.10740*.

Boito, M. Z., Villavicencio, A., and Besacier, L. (2019a). Empirical evaluation of sequence-to-sequence models for word discovery in low-resource settings. *arXiv preprint arXiv:1907.00184*.

Boito, M. Z., Villavicencio, A., and Besacier, L. (2019b). Leveraging translations in multiple languages for low-resource unsupervised word segmentation. Unpublished work. Paper under review.

Cotterell, R., Mielke, S. J., Eisner, J., and Roark, B. (2018). Are all languages equally hard to language-model? *arXiv preprint arXiv:1806.03743*.

Evans, N. and Sasse, H.-J. (2004). Searching for meaning in the library of babel: field semantics and problems of digital archiving. Open Conference Systems, University of Sydney, Faculty of Arts.

Godard, P., Adda, G., Adda-Decker, M., Benjumea, J., Besacier, L., Cooper-Leavitt, J., Kouarata, G.-N., Lamel, L., Maynard, H., Müller, M., et al. (2017). A very low resource language speech corpus for computational language documentation experiments. *arXiv preprint arXiv:1710.03501*.

Godard, P., Zanon Boito, M., Ondel, L., Berard, A., Yvon, F., Villavicencio, A., and Besacier, L. (2018). Unsupervised word segmentation from speech with attention. In *Interspeech*.

Haspelmath, M. (2011). The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia linguistica*, 45(1):31–80.

Pourquoi se tourner vers le SUD
L'importance de choisir un schéma d'annotation en dépendance surface-syntaxique

Kim Gerdes[1] Bruno Guillaume[2]
Sylvain Kahane[3] Guy Perrier[2]
(1) Sorbonne Nouvelle, LPP (CNRS), Almanach (Inria)
(2) Université de Lorraine, CNRS, Inria, LORIA, Nancy
(3) Université Paris Nanterre, Modyco (CNRS)
kim@gerdes.fr, bruno.guillaume@inria.fr
sylvain@kahane.fr, guy.perrier@loria.fr

RÉSUMÉ

L'article défend le schéma d'annotation Surface-Syntactic Universal Dependencies (SUD) comme alternative au schéma standard des Universal Dependencies (UD) pour des projets d'annotation syntaxique, en particulier sur des textes oraux ou non-standard, menés dans un objectif comparatif et typologique.

ABSTRACT

Why you should turn SUD – The importance of choosing a Surface-Syntactic dependency annotation scheme. The article attempts to promote the Surface-Syntactic Universal Dependencies (SUD) annotation scheme to syntactic annotation projects, as an alternative to the standard Universal Dependencies (UD) scheme, particularly on oral or non-standard texts, conducted for comparative and typological studies.

MOTS-CLÉS : corpus arboré, treebank, syntaxe, dépendance, schéma d'annotation, critères distributionnels, tête fonctionnelle.

KEYWORDS: treebank, syntax, dependency, annotation scheme, distributional criteria, functional head.

1 Introduction

Reste-t-il toujours d'actualité d'a rmer que tous les schémas d'annotation syntaxique se valent, avec certains avantages et certains inconvénients ? Ces dernières années, avec l'expérience accumulée du développement de centaines de corpus arborés dans le monde, la question n'est toujours pas

close mais il y a du progrès. Par exemple l'annotation en constituants n'est plus à l'ordre du jour, en premier lieu pour des raisons d'e cacité du processus d'annotation ; les constituants, si jamais ils restaient nécessaires pour une analyse spéci que, peuvent être calculés automatiquement (et de manière plus cohérente) à partir d'un arbre de dépendance. Avec les dépendances, il convient de penser le choix de l'annotation, même si, avec des outils de transformation de graphes tel que Grew (Bonfante et al. 2018), il est possible de convertir un corpus annoté d'un schéma dans un autre. Nous présentons ici quelques usages de corpus arborés et discutons brièvement sur quelle base il est possible de comparer deux schémas d'annotation en dépendance. Nous présentons ensuite le schéma des Surface-syntactic Universal Dependencies (SUD, Gerdes et al. 2018, 2019, https://surfacesyntacticud.github.io/) et montrons quels sont ses atouts.

2 Critères de distinction entre schéma d'annotation

2.1 À quoi sert un corpus arboré ?

Jusqu'à aujourd'hui, une grande partie des corpus arborés sont développés dans un but applicatif : entraîner un analyseur qui, lui, sert indirectement dans un processus TAL de compréhension de textes. L'exploitation proprement linguistique d'un treebank en est encore à ses balbutiements. On peut lister comme buts du développement d'un treebank :

1. Exploiter avec un système TAL
2. Tester un schéma d'annotation théorique
3. Découvrir ou véri er des tendances d'usage de constructions syntaxiques
4. Comparer l'usage de structures syntaxiques entre di érentes langues

Au point 1, il convient d'ajouter que les systèmes statistiques et neuronaux actuels ont avant tout besoin d'une grande masse de données la plus cohérente possible. Ainsi, pour entraîner un système, il est souvent préférable d'encoder moins d'information de manière rapide et cohérente plutôt qu'une information plus riche et plus di cile à mettre sur les données linguistiques.

L'intégration d'une analyse théorique dans un guide d'annotation permet de tester à quel point l'analyse est réellement opérationnelle. Mais ceux qui s'avancent dans cette direction empirique se heurtent souvent à l'omniprésence de phénomènes syntaxiques qui d'une part n'ont aucun lien avec leur problème et d'autre part pour lesquels ils n'y a pas d'analyse facile et généralement acceptée (noms propres, entités nommées, expressions gées, verbes supports, dates, titres d'œuvres, mots étrangers, interjections, reformulations, phrases agrammaticales…). Pour de telles véri cations d'une analyse théorique sur corpus, il convient donc d'intégrer son analyse dans un schéma existant (éventuellement en le modi ant légèrement) plutôt que de commencer des analyses sur des phrases nues. Il est ainsi important que le schéma soit facile d'accès et proche des structures habituellement considérée en syntaxe.

Le projet Universal Dependencies (UD, De Marne e et al. 2014, Nivre et al. 2019, universaldependencies.org) prévoit l'intégration d'analyses idiosyncratiques plus nes en proposant un jeu invariable de relations syntaxiques pour toutes les langues, mais ces relations principales peuvent être ra nées par des relations secondaires, séparées par deux points

(*relationPrincipale:relationSecondaire*). Similairement, le jeu des parties du discours est clos, mais des traits morpho-syntactiques peuvent être ajoutés à volonté à chaque token.
Par contre, la primauté des mots lexicaux dans l'arbre syntaxique UD amène à des structures très inhabituelles comme par exemple l'analyse des prépositions en tant que dépendant « casuel » du nom (appelée aussi l'analyse turque de l'anglais). En plus les analyse UD résultent dans des structures similaires entre langues même si, structurellement, les langues divergent dans la réalisation d'une construction. Ainsi, certaines mesures des di érences typologiques ne sont pas possibles sur UD directement et nécessitent une transformation des treebanks (Gerdes et Kahane 2016, Osborne et Gerdes 2019).

2.2 Comment choisir un schéma d'annotation ?

On peut retenir les critères suivants. Le schéma doit

1. se baser sur des critères syntaxiques (et non sémantiques) si on veut :
   a. l'appliquer à des langues typologiquement di érentes ;
   b. mesurer des di érences syntaxiques entre langues ;
2. faciliter l'annotation par des critères distributionnels que l'annotateur peut appliquer de manière reproductible sans recourir à des lexiques extérieurs ;
3. distinguer d'une part une grille d'analyse obligatoire et universelle et d'autre part permettre des sous-spéci cations et des ra nements idiosyncratiques des analyses (par langue ou par treebank) ;
4. s'intégrer dans les projets internationaux de développement de treebanks ;
5. se rapprocher des analyses classiques a n de faciliter des requêtes dans le treebank et des extensions du schéma ;
6. se limiter à un système de traits par tokens et de relations de dépendances hiérarchiques (c'est-à-dire un nœud domine l'autre) et binaires entre tokens, même si toutes les relations ne rentrent pas parfaitement dans ce schéma (e.g. les coordinations).

Le dernier point permet l'utilisation d'outils de visualisation, d'annotation et d'analyse automatique, – essentielle pour un processus d'annotation avec des boucles de bootstrapping.

3 Surface-syntactic Universal Dependencies

Le schéma Surface-syntactic Universal Dependencies (SUD, Gerdes et al. 2018, 2019, https://surfacesyntacticud.github.io/) est le résultat de l'expérience accumulée dans le développement de corpus arborés dans plusieurs projets (ANR Rhapsodie, ANR Orféo, ANR NaijaSynCor, ANR Pro térole, projet Procore « Semi-automatic Creation of a Parallel Treebank of Cantonese and Mandarin »). SUD se fonde sur des critères distributionnels et suit ainsi l'analyse dépendentielle classique (Hudson 1984, 1987, Mel'cuk 1988, Prague Dependency Treebank, Hajič & Hajičová 1997) avec des têtes fonctionnelles.
SUD est presque isomorphe à UD, dans le sens qu'une annotation SUD peut être transformée en annotation UD et vice versa avec peu de perte – les pertes peuvent principalement être attribuées à des analyses non-conformes avec les guides UD ou SUD (mais il y a aussi des pertes causées par la structure UD même, qui est plus plate que la structure SUD et ne contient pas tous les liens hiérarchiques entre dépendants) La transformation des treebanks UD en SUD permet de corriger

quelques liens problématiques pour les mesures comparatives et facilite ainsi des études typologiques sur les treebank UD.

Toutes les parties du discours et une grande partie des relations (boîte orange de la Figure 1 ci-dessous) sont les mêmes en UD et en SUD. Les deux schémas se distinguent avant tout dans l'analyse des arguments verbaux et prépositionnels (boîte bleue ci-dessous, des liens très fréquents dans les corpus). Les relations spéci quement SUD sont disposées dans une taxonomie, ce qui permet la sous-spéci cation d'une relation si une construction ne permet pas de choisir entre deux relations. SUD considère principalement 3 types de dépendants verbaux : sujet (*subj*), modi eurs (*mod*) et complément (*comp*). Les compléments peuvent être di érenciés en cinq types : Les arguments obliques (*comp:obl*), les arguments directs (*comp:obj*) incluant l'argument d'une préposition, le lien entre auxiliaire et verbe lexical (*comp:aux*), le lien entre tête d'une clivée et le noyau de la phrase (*comp:cleft*) et nalement les attributs (*comp:pred*). Toute information référant à la relation sémantique entre deux unités est clairement séparée de la syntaxe mais peut optionnellement être ajoutée aux relations, séparées par le symbole arobase (@, boîtes bleues-claires ci-dessous).

Le schéma SUD a été développé et appliqué dans le contexte du développement de corpus non-standard, en particulier de l'oral. Des guides d'annotation SUD ont été développés pour le français, le naija et le chinois, démontrant ainsi la versatilité de l'approche SUD à l'annotation syntaxique.

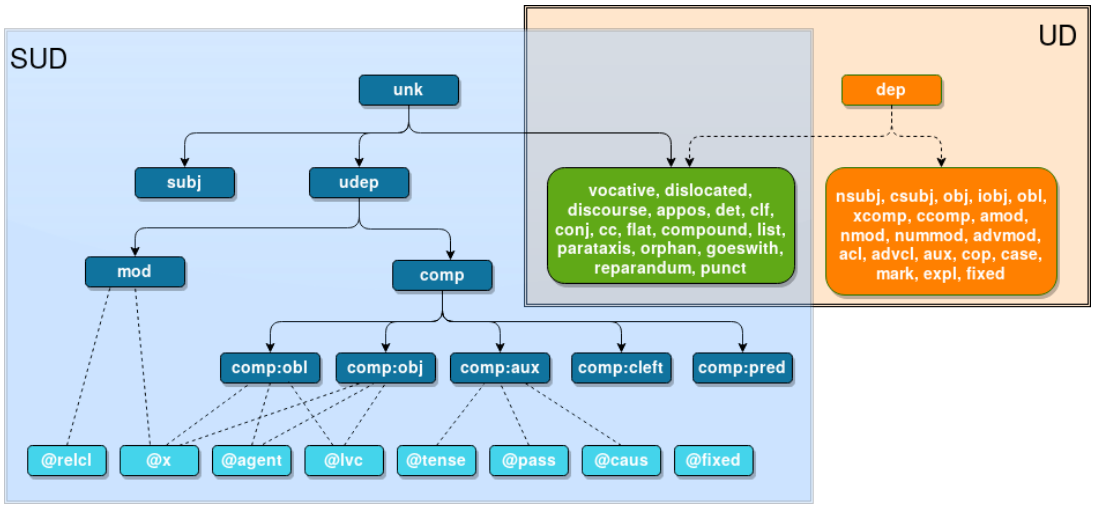


Figure 1 : Le schéma des relations SUD

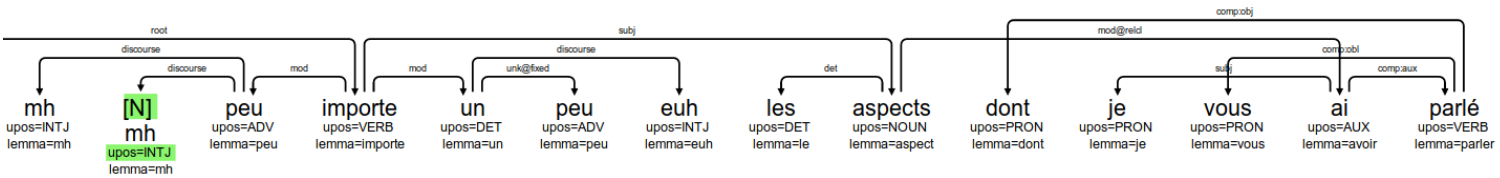


Figure 2 : exemple français d'une analyse en SUD

Références

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

De Marne e, M. C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., & Manning, C. D. (2014). *Universal Stanford dependencies: A cross-linguistic typology.* In *LREC* (Vol. 14, pp. 4585-4592).

Gerdes, K. and Kahane, S. (2016). Dependency annotation choices: Assessing theoretical and practical issues of universal dependencies. In *Proceedings of the 10th Linguistic Annotation Workshop* held in conjunction with ACL 2016 (LAW-X 2016) (pp. 131-140).

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

Gerdes, K., Guillaume, B., Kahane, S., & Perrier, G. (2019). Improving Surface-syntactic Universal Dependencies (SUD): surface-syntactic relations and deep syntactic features. *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT 2019)* held at the Syntaxfest 2019.

Hajič, J., & Hajičová, E. (1997). Syntactic tagging in the prague tree bank. In Proceedings of the Second European Seminar "Language Applications for a Multilingual Europe (pp. 55-68).

Hudson, R.A. (1984). *Word grammar.* Oxford: Blackwell.

Hudson, R.A. (1987). Zwicky on heads. *Journal of linguistics,* 23(1), pp.109-132.

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

Nivre J. et al. (2019). *Universal dependencies 2.4*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Osborne, T., Gerdes K. (2019). The status of function words in dependency grammar: A critique of Universal Dependencies (UD). *Glossa: a journal of general linguistics*, 4(1).

Advancing the study of endangered languages with computational tools for morphology: The case of Asama verb paradigms

RÉSUMÉ

Bien que de nombreux outils informatiques soient conçus pour des langues bien dotées, ceux-ci peuvent néanmoins également bénéficier aux langues sous-documentées et en danger. Ceci peut être illustré par notre projet en cours de description et de documentation de l'asama, une langue en danger du Japon, utilisant des transducteurs à états finis pour implémenter une grammaire de la morphologie verbale de cette langue. Cette implémentation nous a permis de vérifier directement l'exactitude de notre analyse, de corriger les erreurs, et de combler ses lacunes. Ces outils ont aussi été mis à contribution pour améliorer et accélérer la production de textes à glose interlinéaire avec les outils standards des linguistes de terrain. Ces outils fournissent également un moyen simple d'explorer de manière quantitative la complexité des systèmes morphologiques.

ABSTRACT

Advancing the study of endangered languages with computational tools for morphology: The case of Asama verb paradigms

Though many computational tools are designed for high-resource languages, they can also benefit under-described and endangered languages. This is exemplified by our ongoing project on the description and documentation of Asama, an endangered language of Japan. We focus on the use of finite-state transducers to implement a grammar of verb morphology. This computational implementation enabled us to directly check the accuracy of our analysis, correct mistakes, and fill gaps in the description. These tools were also put to use for improving and speeding up the production of interlinar glossed texts with standard tools used by fieldworkers. Such tools also provide a straightforward mean to quantitatively explore the complexity of morphological systems.

MOTS-CLÉS : langues en danger, morphologie, transducteurs à états finis, textes avec glose interlinéaire, complexité morphologique, langues japoniques.

KEYWORDS: endangered languages, morphology, finite-state transducers, interlinear glossed texts, morphological complexity, Japonic languages.

1 Introduction

Asama is a highly endangered and under-described Japonic language spoken on the Tokunoshima Island, Japan. As part of the project of writing a reference grammar of Asama, computational tools were used to develop an implemented description of the inflectional morphology of Asama. The goal of this computational implementation was threefold: 1. to test the accuracy and coverage of the analysis; 2. to partially automate the glossing of texts; 3. to enable quantitative measures of morphological complexity.

2 Finite-state transducers and linguistic analysis

Though most Japonic languages exhibit a rather simple and transparent verb morphology close to the canonical agglutinative type, Asama displays several non-canonical phenomenon, like alternation of tones and vowel length in inflected forms (Table 1). These are difficult to handle in a straightforward and satisfying manner with morpheme-based models and are better accounted for by a Word-and-Paradigm approach (Matthews, 1972; Stump, 2001; Blevins, 2016).

Table 1: Stem and tone alternations in Asama's verb morphology

Inflection	TUBJUI "to fly"	KOORAKJUI "to dry"	UCJUI "to hit"
NPST	*tubj-* H	*koorakj-* H	*ucj-* LH
PROH	*tubj-* H	*koorakj-* H	*ucj-* LH
NEG	*tub-* H	*koorak-* H	*ut-* LH
IMP.INDIR	*tub-* H	*kooraak-* H	*uut-* HL
PST	*tud-* H	*kooracj-* H	*uccj*-LH
SEQ	*tud-* H	*kooraacj-* H	*uccj*-HL
PROG	*tud-* LHL	*kooracj-*, *kooraacj-* H	*uccj*-HL

Convinced of "the insufficiency of paper-and-pencil linguistics" (Karttunen, 2006) and of both the advantages and the feasability of developing implemented morphological descriptions for endangered languages (Jacques et al., 2012; Snoek et al., 2014; Pellard and Yamada, 2017), we concluded it was desirable to use computational tools in order to provide a maximally precise and explicit description of Asama verb morphology.

Finite-state transducers have several desirable properties for our purpose: they have sufficient power to handle (almost) any morphological phenomena (Beesley and Karttunen, 2003), they can be used to implement various theoretical frameworks (Karttunen, 2003), and they are reversible, i.e. they can be used for both production and recognition of inflected forms. We used the Foma free software (Hulden, 2009), whose syntax is similar to that of phonological rewrite rules familiar to many linguists.

From a single morphological grammar, three different transducers were implemented as mappings 1. from a content paradigm cell $\langle L, \sigma \rangle$ pairing a verbal lexeme L with a set of morphosyntactic properties σ to the form w realizing that cell, 2. from a lexeme L to its full realized paradigm (a set of cells $\{\langle w, \sigma \rangle, \langle w', \sigma' \rangle, \ldots\}$), 3. from an unlabeled form w to its different possible morphological analyses.

These transducers were helpful for testing hypotheses, e.g. about inflectional classes, and for verifying the accuracy of the description, i.e. by checking that correct forms are obtained and that there is no overgeneration. They also facilitated elicitation of verb paradigms during fieldwork.

3 Finite-state transducers and linguistic documentation

Compared to most computational linguistics projects, research on endangered languages usually relies on small corpora developed by a single linguist or a small team, hand in hand with grammatical description. Many natural language processing tools of automatic parsing or machine learning would thus be difficult to use in the context of low-resource endangered languages. Field linguists have thus

relied on less sophisticated tools for creating corpora of interlinear glossed texts. The most widely used tools among fieldworkers are the Field Linguist's Toolbox and FieldWorks Language Explorer (FLEx). However, the morphological parser of these softwares impose a morpheme-based item-and-arrangement approach where words are segmented into linear sequences of morphemes whose underlying forms are listed in a lexical database. This is problematic for linguists with other theoretical views on morphology, but more crucially, it creates difficulties for handling non-concatenative phenomena, such as the vowel length and tone alternations of Asama.

Finite-state transducers provide a good compromise for field linguists since they do not enforce a particular theoretical approach to morphology, allow to deal with non-concatenative phenomena, use a format familiar to most linguists, and have an acceptable learning curve. Finite-state transducers thus allowed us to produce complete paradigms for all known lexemes and from there all $\{w, \text{gloss}\}$ pairings. It was then trivial with a script to apply a transducer to each verb in a Toolbox lexical database and to export the results as a database of inflected forms that can be used by the parser of Toolbox.This method both facilitates and speeds up the glossing procedure, which allows to easily produce a larger quantity of interlinear glossed texts and thus helps advancing language documentation.

4 Finite-state transducers and theoretical morphology

Recent work (Ackerman et al., 2009; Ackerman and Malouf, 2013; Bonami and Luís, 2015) in theoretical morphology has shown how information theory could be used to measure the complexity of morphological systems, in particular the implicative structure of inflectional paradigms, and thus solve the *Paradigm Cell Filling Problem* ("What licenses reliable inferences about the inflected (and derived) surface forms of a lexical item?", Ackerman et al., 2009, 54). The implemented transducers were composed together in order to obtain for any given realized cell $\langle w, \sigma \rangle$ the list of all possible realized forms of any other paradigm cell. It was then possible to quantify the amount of uncertainty in predicting an unknown inflectional form based on another known form as a conditional entropy measure (Table 2).

Table 2: Conditional entropy between paradigm cells (unary implications) measured in bits

$H(C\|L)$	NPST	CONV	IMP.INDIR	SEQ	PROG.NPST
NPST		0.000	0.000	0.222	0.244
CONV	0.923		0.082	0.296	0.550
IMP.INDIR	0.950	0.211		0.222	0.525
SEQ	1.362	1.012	0.425		0.317
PROG.NPST	1.204	0.910	0.423	0.000	

These calculations enabled us to identify the different loci of uncertainty and complexity within the system. This confirmed several of our preliminary hypotheses about the role of the neutralization of vowel length and tone distinctions and about the opacity of some segmental alternations. This also enabled us to identify the principal parts of the system (Stump and Finkel, 2013, 11), the minimal set of forms needed to predict a lexeme's complete paradigm, which allowed to elicit only those forms instead of full paradigms when time was missing during fieldwork.

References

Ackerman, F., Blevins, J. P., and Malouf, R. (2009). Parts and wholes. In Blevins, J. P. and Blevins, J., editors, *Analogy in grammar*, pages 54–81. Oxford University Press, Oxford.

Ackerman, F. and Malouf, R. (2013). Morphological organization. *Language*, 89(3):429–464.

Beesley, K. and Karttunen, L. (2003). *Finite state morphology*. Center for the Study of Language and Information Publications, Stanford.

Blevins, J. P. (2016). *Word and paradigm morphology*. Oxford University Press, Oxford.

Bonami, O. and Luís, A. R. (2015). Sur la morphologie implicative dans la conjugaison du portugais. In Léonard, J. L., editor, *Morphologie flexionnelle et dialectologie romane*, pages 111–151. Peeters, Leuven.

Hulden, M. (2009). Foma. In *Proceedings of the Demonstrations Session at EACL 2009*, pages 29–32. Association for Computational Linguistics.

Jacques, G., Lahaussois, A., Michailovsky, B., and Rai, D. B. (2012). An overview of Khaling verbal morphology. *Language and Linguistics*, 13(6):1095–1170.

Karttunen, L. (2003). Computing with realizational morphology. In Gelbukh, A., editor, *Computational linguistics and intelligent text processing*, pages 203–214. Springer, Berlin.

Karttunen, L. (2006). The insufficiency of paper-and-pencil linguistics. In Butt, M., Dalrymple, M., and Holloway King, T., editors, *Intelligent linguistic architectures*, pages 287–300. Center for the Study of Language and Information Publications, Stanford.

Matthews, P. H. (1972). *Morphology*. Cambridge University Press, Cambridge, 1 edition.

Pellard, T. and Yamada, M. (2017). Verb morphology and conjugation classes in Dunan (Yonaguni). In Kiefer, F., Blevins, J. P., and Bartos, H., editors, *Perspectives on morphological organization*, pages 31–49. Brill, Leiden.

Snoek, C., Thunder, D., Lõo, K., Arppe, A., Lachler, J., Moshagen, S., and Trosterud, T. (2014). Modeling the noun morphology of Plains Cree. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 34–42.

Stump, G. and Finkel, R. A. (2013). *Morphological typology*. Cambridge University Press, Cambridge.

Stump, G. T. (2001). *Inflectional morphology*. Cambridge University Press, Cambridge.

Liste des sponsors

Groupement de recherche "Linguistique Informatique, Formelle et de Terrain"
La linguistique informatique met à la disposition des linguistes un large éventail de techniques et de ressources. L'objectif du Groupement de recherche "Linguistique Informatique, Formelle et de Terrain" est d'explorer ce potentiel en tirant parti d'un réseau scientifique favorisant les interactions entre linguistes, linguistes de terrain et linguistes informaticiens et de favoriser l'émergence de méthodes nouvelles.

Institut des Sciences Humaines et Sociales du CNRS
Le Groupement de Recherche "Linguistique Informatique, Formelle et de Terrain" relève de l'Institut des Sciences Humaines et Sociales du CNRS.

Laboratoire Ligérien de Linguistique
Le Laboratoire Ligérien de Linguistique est une Unité Mixte de Recherche (UMR 7270) associant au CNRS les linguistes des universités d'Orléans et de Tours et des conservateurs et agents du Département de l'Audiovisuel de la Bibliothèque nationale de France (BnF).

Langues et Civilisations à Tradition Orale

Le LACITO est un laboratoire de recherche pluridisciplinaire (linguistique et anthropologie) qui se consacre prioritairement à l'étude des langues à tradition orale.

INALCO

L'Institut national des langues et civilisations orientales (INALCO), dit Langues O', est un établissement français d'enseignement supérieur et de recherche qui enseigne les langues et les civilisations depuis 1795. Enseignement et recherche à l'INALCO concernent les langues et civilisations autres que celles originaires d'Europe occidentale.

Université Sorbonne Nouvelle - Paris 3

Les Sciences du langage sont l'une des principales spécialités de l'Université Sorbonne Nouvelle - Paris 3. La Sorbonne Nouvelle est notamment partenaire des centres de recherche suivants : UMR 7018 Laboratoire de Phonétique et de Phonologie (LPP), UMR 7107 Langues et Civilisations à Tradition Orale (LACITO), UMR 7597 Histoire des Théories Linguistiques (HTL), FRE 2018 Monde Iranien et Indien et UMR 8094 Langues, Textes, Traitements Informatiques, Cognition (LaTTiCe).

Liste des auteurs