Actes des 5èmes journées du Groupement de Recherche CNRS “ Linguistique Informatique, Formelle et de Terrain ”

Karën Fort, Claire Gardent, Yannick Parmentier

5èmes journées du Groupement de Recherche CNRS « Linguistique Informatique, Formelle et de Terrain »

LIFT 2023

https://lift2023.sciencesconf.org

Karën Fort, Claire Gardent, Yannick Parmentier (Éds.)

Nancy, France, 20-21 novembre 2023

Préface

La linguistique informatique met à la disposition des linguistes un large éventail de techniques et de ressources qui ouvrent des perspectives nouvelles pour l'analyse linguistique, que ce soit pour collecter et annoter des données ou pour extraire ou vérifier des généralisations linguistiques. L'objectif du Groupement de Recherche LIFT (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 informaticien.ne.s et de favoriser l'émergence de méthodes nouvelles qui bénéficient à la fois aux linguistes (automatisation des processus d'analyse et de validation), aux linguistes de terrain (facilitation des processus de collecte et d'analyse des données) et aux linguistes informaticien.ne.s (développement de nouvelles techniques nécessitées par l'analyse linguistique, essor des méthodes non ou faiblement supervisées pour l'analyse des langues peu dotées, peu écrites ou non documentées).

Les journées LIFT sont un rendez-vous annuel qui offre un forum d'échanges aux trois communautés de linguistes. Cet événement, qui accueille habituellement une petite cinquantaine de participants, couvre des sujets très divers tels que les approches formelles, la documentation linguistique ou encore le traitement automatique des dialectes et des langues peu dotées.

Cette année, 18 communications seront présentées sous formes de posters et deux panels interactifs d'1h30 seront organisés autour des questions "Approches fondées sur l'expertise vs approches fondées sur les données : quelle place pour la linguistique ?", animé par Yannick Parmentier de l'Université de Lorraine, et "Variation et diversité linguistique : quelles questions se posent encore face à la variabilité dans les données langagières ?", animé par Delphine Bernhard de l'Université de Strasbourg. L'alternance de sessions posters avec ces panels de discussion devraient permettre de susciter des échanges fructueux.
Les trois conférencièr.e.s invité.e.s présenteront des travaux illustrant chacun un aspect différent des thèmes LIFT. François Yvon (ISIR, Sorbonne Université & CNRS) adressera le thème du multilinguisme avec une présentation intitulée "Construire et évaluer des modèles de langue massivement multilingues". Sylvain Kahane (Modyco, Université Paris Nanterre & CNRS / Institut Universitaire de France) parlera de documentation linguistique avec un exposé intitulé "Règles de grammaires et corpus annotés - Autour du projet Autogramm". Enfin, la présentation de Agata Savary (LISN, Université Paris Saclay & CNRS) intitulée "Nous croyions que les yeux de la coréférence étaient fermés sur les expressions polylexicales et ils le sont la plupart de temps" portera sur le phénomène linguistique des expressions polylexicales.

Ces journées 2023 sonnent le glas de LIFT 1. Un second épisode est d'ors et déja validé par le CNRS avec pour nouvelle porteuse, Karën Fort (LORIA, Sorbonne Université).

Bonne conférence et bonne lecture !
Karën, Claire et Yannick Nancy, le 17 novembre 2023

Comités

Comité d'organisation

— Marie Baron
— Karën Fort
— Nathalie Fritz
— Claire Gardent
— Sylvie Hilbert
— Anne Marie Messaoudi
— Yannick Parmentier

Comité de programme

— Maxime Amblard Université de Lorraine
— Delphine Bernhard LiLPa, Université de Strasbourg
— Hee-Soo Choi LORIA - ATILF
— Berthold Crysmann CNRS - LLF (UMR 7110) - U Paris
— Fanny Ducel Sorbonne Université
— Solène Evain Laboratoire d'Informatique de Grenoble
— Karën Fort Sorbonne Université
— Claire Gardent CNRS/LORIA Nancy
— Kim Gerdes LISN, CNRS and University Paris-Saclay
— Sylvain Kahane Modyco, Université Paris Ouest Nanterre & CNRS
— Anaïs Lefeuvre-Halftermeyer LIFO - Université d'Orléans
— Sylvain Loiseau UMR ICAR
— Alexis Michaud CNRS - LACITO (Langues et Civilisations à Tradition Orale, Centre National de la Recherche Scientifique)
— Aurélie Névéol Université Paris-Saclay, CNRS, LISN
— Yannick Parmentier LORIA - Université de Lorraine
— Thierry Poibeau LaTTiCe-CNRS
— Emmanuel Schang LLL, Univ. Orléans & CNRS
— Guillaume Wisniewski LLF - Université de Paris Cité

Présentations invitées

Sylvain Kahane (Modyco, Université Paris Nanterre & CNRS / Institut Universitaire de France)

Titre : Règles de grammaires et corpus annotés - Autour du projet Autogramm

Résumé : Dans cette présentation, nous discuterons de ce qu'est une règle de grammaire et de la façon dont on peut extraire de telles règles d'un corpus annoté. Ce questionnement est au centre du projet ANR Autogramm (Modyco, Lacito, Lisn, Loria-Sémagram), sur l'induction de grammaires descriptives à partir de corpus annotés. Nous insisterons sur l'intérêt d'avoir des règles de grammaire quantifiées et ordonnées pour la caractérisation d'un corpus et à travers lui d'une langue ou d'un état de langue. Nous présenterons les différents travaux éffectuées dans le cadre du projet concernant le développement de treebanks, d'outils d'annotation et d'extraction automatique de règles de grammaire.

Agata Savary (LISN, Université Paris Saclay & CNRS)

Titre : Nous croyions que les yeux de la coréférence étaient fermés sur les expressions polylexicales et ils le sont la plupart de temps

Résumé : Les expressions polylexicales sont des combinaisons de plusieurs mots qui possèdent des propriétés sémantiques particulières, comme des degrés variés de compositionnalité sémantique, la décomposabilité, la transparence et la figuration. Plusieurs débats linguistiques suggèrent que les idiosyncrasies sémantiques de ces types conditionnent les configurations morpho-syntaxiques dans lesquelles une expressions polylexicale donnée peut apparaître. Nous étendons cette argumentation à la coréférence nominale. Nous posons l'hypothèse que les composants internes d'une expression polylexicale sont peu susceptibles d'appartenir à des chaînes coréférentielles. Bien que des travaux antérieurs aient remarqué la rareté des phénomènes liés à la coréférence en présence d'expressions polylexicales, à notre connaissance, cette observation n'avait pas été quantifiée. Nous comblons cette lacune par une étude des intersections entre les expressions polylexicales verbales et la coréférence nominale dans des corpus français. Les résultats corroborent largement notre hypothèse mais montrent également des tendances variables selon les types d'expressions polylexicales et le genre du corpus. L'analyse des certains exemples révèle des propriétés intéressantes de la coréférence, notamment dans en parole spontanée.

François Yvon (ISIR, Sorbonne Université & CNRS)

Titre : Construire et évaluer des modèles de langue massivement multilingues

Résumé : Dans cette présentation, je discuterai des difficultés que pose l'apprentissage et l'évaluation de modèles de langue massivement multilingues, capables de prendre en charges des dizaines, voire des centaines de langues. Après avoir motivé l'apprentissage de tels modèles, je m'arrêterai, en m'appuyant sur l'expérience du développement du modèle Glot-500 et des ressources associées, sur la question de la prise en charge de langues "moins bien dotées", c'est-à-dire pour lequelles les données d'apprentissage sont souvent lacunaires, très spécialisées, ainsi que possiblement très bruitées.

Table des matières

Première partie

Session dédiée aux langues peu dotées

Actes des journées LIFT 2023
Nancy, France, 20-21 novembre 2023

Outiller la documentation des langues créoles

Éric Le Ferrand[1,2], Claudel Pierre-Louis[1], Ruoran Dong[3], Benjamin Lecouteux[3], Daphne Gonçalves-Teixeira[1], William N. Havard[1], Emmanuel Schang[1]
(1) LLL, 10 rue de Tours, BP 46527 - 45072 ORLEANS CEDEX 2 (FRANCE)
(2) Boston College, Boston (USA)
(3) LIG, Université Grenoble Alpes (FRANCE)
william.havard@univ-orleans.fr, leferran@bc.edu,
daphne.goncalves-teixeira@univ-orleans.fr,
benjamin.lecouteux@univ-grenoble-alpes.fr,
emmanuel.schang@univ-orleans.fr

RÉSUMÉ

Ce papier propose donc un retour d'expérience basé sur l'emploi d'outils informatiques sur différents terrains linguistiques concernant les langues créoles.

ABSTRACT

Tooling up Creole Languages Documentation

This paper provides feedback based on the use of IT tools in different linguistic fields concerning Creole languages.

MOTS-CLÉS : Documentation des langues, créoles, traitement automatique de la parole.

KEYWORDS: Language Documentation, Creole Languages, Automatic Speech Processing.

1 Introduction

Dans ce papier, nous passons en revue les pistes suivies par le projet CREAM (documentation des langues CREoles Assistée par la Machine, ANR CS38) pour la documentation outillée des langues créoles. Afin de répondre aux besoins distincts de ses utilisateurs, principalement des linguistes, plusieurs solutions sont présentées.

La documentation des langues consiste à mettre à disposition des données langagières sous une forme consultable par ses utilisateurs [1].

> Documentation of a language is an activity (and, derivatively, its result) that gathers, processes and exhibits a sample of data of the language that is representative of its linguistic structure and gives a fair impression of how and for what purposes the language is used. (Lehmann, 2001)

1. Voir également : (Austin, 2016; Michaud *et al.*, 2016) entre autres.

Il est essentiel de noter que les langues créoles ne partagent pas les mêmes niveaux d'écriture et de standardisation. Certaines langues jouissent d'un statut officiel, à l'instar du santoméen à São Tomé et Príncipe, l'une des langues nationales de l'archipel. D'autres, comme le créole de l'île Maurice, sont parlées par la majorité de la population sans bénéficier d'un statut officiel reconnu.

Alors qu'il est généralement admis que le Traitement Automatique des Langues (TAL) vise principalement à faciliter la graphisation (Chaudenson, 2005) des langues minoritaires, notre approche diverge en ne conférant pas à la graphisation un rôle prédominant au sein de ce projet. En effet, les résultats de nos travaux sur le terrain nous orientent vers l'utilisation des outils du TAL à des fins autres que la transcription.

Ainsi, cet article apporte un retour d'expérience sur l'application d'outils informatiques dans divers contextes linguistiques en milieu créolophone.

2 Transcription automatisée

Le besoin de transcription automatique et les modalités de sa réalisation sont intrinsèquement liés à l'existence d'une norme orthographique stable et acceptée. Pour les langues créoles qui disposent d'une orthographe normalisée, comme le créole de la Guadeloupe (kréyòl gwadloupéyen), la transcription automatique est d'une aide précieuse pour le linguiste. Nous avons reçu plusieurs demandes de transcription automatique émanant de linguistes locaux ou de projets de recherche récoltant des données de terrain.

2.1 Transcrire automatiquement depuis le terrain

Au cours d'un travail de terrain en Guadeloupe effectué par l'un des auteurs en soutien à un projet de recherche (NSF-IRES 1952568 : Experimental linguistics in the Caribbean), un ensemble d'enregistrements de terrain nous a été fourni avec pour seule directive de fournir des transcriptions générées automatiquement. Ces enregistrements présentaient une grande diversité en termes de nature et de qualité. Certains comprenaient une série de jugements grammaticaux en français, tandis que d'autres contenaient des discussions en anglais sur des phénomènes linguistiques. La plupart d'entre eux, cependant, renfermaient de la parole spontanée en créole dans divers contextes tels que des visites guidées, des séminaires et des conversations.

Notre sélection a porté sur un enregistrement d'une heure renfermant une quantité suffisante de contenu en créole mêlé à des segments entièrement en français, des segments où la syntaxe créole et française se mélangent (*code-switching*) et une poignée de segments en anglais. Dans un premier temps, nous avons appliqué un algorithme de détection d'activité vocale

(VAD) [2] pour identifier les segments de parole. À cette fin, nous disposions d'un modèle de reconnaissance de la parole préalablement entraîné sur une heure de parole transcrite (Macaire *et al.*, 2022).

Ce modèle repose sur l'architecture WAV2VEC2 (Baevski *et al.*, 2020), une architecture de transformers utilisant un apprentissage auto-supervisé pour extraire les paramètres acoustiques. Dans notre cas spécifique, le modèle LeBenchmark a été employé (Evain *et al.*, 2021). La représentation générée par cette architecture est ensuite alimentée à une tête de transformers entraînée sur un corpus de parole transcrite, utilisant une fonction CTC (*Connectionist Temporal Classification*, Graves *et al.* 2006). Cette fonction CTC apprend à générer la probabilité d'un caractère pour chaque paramètre acoustique.

Le processus se poursuit avec un décodage final modulé à l'aide d'un modèle de langue, lui-même entraîné sur les mêmes transcriptions qui ont servi à l'entraînement initial du modèle. Cette approche assure une cohérence et une précision accrues dans la transcription automatique des enregistrements en créole, en tirant parti de la représentation riche des caractéristiques acoustiques acquises par l'architecture WAV2VEC2, tout en affinant le résultat final à l'aide d'un modèle de langue formé sur le corpus spécifique utilisé pour l'entraînement initial.

Ainsi, nous avons appliqué le modèle développé sur les segments détectés par la VAD. Un Gold standard, établi en se basant sur nos propres transcriptions, nous a été fourni, permettant ainsi une évaluation concrète des performances du modèle. Les résultats obtenus révèlent une performance relativement faible avec un taux d'erreur par mot (WER) à 73% et un taux d'erreur par caractère (CER) à 45%. Cependant, il est important de noter que malgré ces limitations, l'utilisation de ce modèle offre aux linguistes de terrain un gain de temps considérable, variable toutefois en fonction de la qualité de l'audio.

Parmi les retours que nous avons recueillis de la part des linguistes de terrain, la détection des segments de parole a été particulièrement saluée pour le temps qu'elle permet d'économiser. De plus, bien que notre évaluation semble indiquer une performance modérée du système de transcription, il faut souligner que la qualité des transcriptions varie considérablement au sein d'un même enregistrement. Une transcription générée peut ainsi présenter une faible qualité pour des segments très bruités, tandis que d'autres nécessiteront uniquement des corrections mineures. Cette hétérogénéité souligne la complexité de la tâche de transcription automatique, tout en mettant en lumière les avantages substantiels que notre approche peut offrir dans le contexte spécifique des langues créoles.

2. https://github.com/amsehili/auditok

Analyse d'erreur	Référence	Transcription automatique
La nasale finale est reconnue comme deux orales	zot matinike gwadloupeyen	zoln patinike gwadloup ee
la référence est en français	deux saisons	deu sezon
erreur de segmentation	zo kay an grante	jo kay angrandte
erreurs de segmentation et de transcription	byen pale de bonda nou kay soukre bonda	mye fame de gonda nou ka ai soucebo

TABLE 1 – Examples de transcriptions

2.2 Aligner les transcriptions sur la parole

Par ailleurs, en appliquant les mêmes techniques, nous avons procédé à l'alignement d'un corpus audio en créole haïtien. Ce corpus, composé de 10 enregistrements audio associés à des transcriptions au format MSWord, présentait la particularité de ne pas comporter de balises temporelles facilitant l'alignement. Pour résoudre cette problématique, nous avons opté pour une transcription automatique de type 'gros grain', générée à partir de 10 minutes de parole préalablement alignée, et traitée ultérieurement à l'aide de WAV2VEC2.

Cette approche a permis d'obtenir une représentation plus robuste du contenu audio, malgré l'absence initiale de balises temporelles. L'utilisation de la transcription automatique 'gros grain' a servi de prétraitement, suivie d'une étape cruciale impliquant le modèle WAV2VEC2 pour affiner et préciser l'alignement temporel. Cette démarche souligne l'adaptabilité de notre méthode dans des contextes divers, renforçant ainsi sa pertinence pour des langues créoles variées et des corpus audio présentant des défis spécifiques.

3 TètKole et l'interprétation

TètKole, signifiant 'ensemble' en créole haïtien, représente un outil conçu dans le cadre du projet CREAM par des étudiants en informatique de l'université de Chambéry. Il s'agit d'un redéveloppement de l'outil LIG-AIKUMA, qui était précédemment disponible sur la plateforme Android (Blachon *et al.*, 2016). TètKole a pour fonction de permettre l'interprétation (traduction à l'oral) d'enregistrements réalisés par des linguistes de terrain. Il vise à l'obtention de corpus de parole parallèles dans un objectif de documentation des langues.

L'utilisation de LIG-AIKUMA dans notre projet s'est heurtée à trois obstacles majeurs, justifiant ainsi le développement de TètKole :

Stabilité sur tablette et téléphone : L'instabilité de son fonctionnement sur tablette ou téléphone a été un point de friction significatif, suscitant des retours négatifs de l'ensemble des utilisateurs.

Préférence des linguistes pour les PC : Les linguistes du projet manifestent une préférence marquée pour travailler sur PC plutôt que sur téléphone ou tablette, constituants initiaux de LIG-AIKUMA. Cette préférence a motivé le besoin d'une version mieux adaptée aux environnements de travail privilégiés par les linguistes.

Limite de durée des fichiers son : Une autre limitation était imposée par la contrainte des téléphones portables et des tablettes, qui ne permettaient pas de travailler sur des fichiers son d'une durée dépassant les 15 minutes. Cette restriction ne répondait pas aux exigences des linguistes du projet, qui avaient exprimé le besoin de manipuler des enregistrements de durées plus étendues.

En réponse à ces limitations, TètKole a été développé pour surmonter ces obstacles spécifiques et répondre de manière plus adéquate aux besoins et préférences des linguistes du projet CREAM.

TètKole est donc un redéveloppement en Java de certaines fonctionnalités de LIG-AIKUMA. L'outil est librement disponible sur https://github.com/LLL-Orleans/TetKole sous licence GPL. Il a été réalisé par des étudiants de Master Informatique de l'université de Savoie Mont Blanc (Chambéry) dans le cadre d'un projet.

TètKole offre une interface conviviale pour les linguistes, leur permettant de charger un enregistrement au format .wav ou .mp3. À l'aide de la souris, le linguiste peut sélectionner avec précision la portion de parole à interpréter, comme illustré par les barres verticales sur la Figure 1. En suivant cette sélection, en cliquant sur l'icône dédiée à l'enregistrement, le linguiste génère un fichier son contenant son interprétation.

Cette approche centrée sur l'interaction visuelle et l'ergonomie vise à simplifier le processus d'interprétation, offrant aux linguistes un moyen intuitif et efficace pour traiter les enregistrements et produire des interprétations orales. La combinaison de la sélection visuelle et de la création rapide de fichiers son facilite le flux de travail des linguistes, favorisant ainsi une utilisation efficace de l'outil TètKole dans le cadre du projet CREAM.

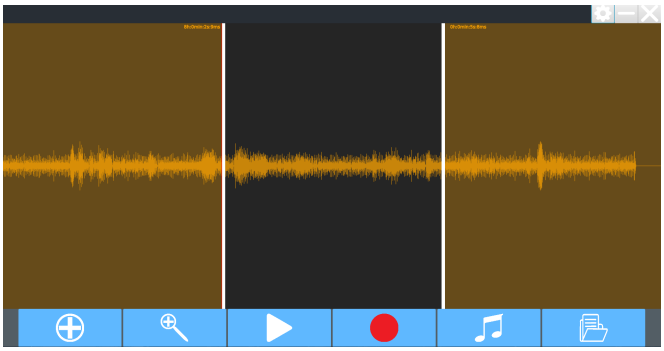


FIGURE 1 – L'écran d'accueil de TètKole après chargement du fichier son à interpréter

Le linguiste peut vérifier la qualité de son travail et reprendre ses interprétations (supprimer les interprétations non satisfaisantes) ou les valider (Figure 2).

Cet outil a permis la réalisation d'un corpus oral parallèle français /haïtien d'environ 3 heures. A partir de la version française, il est aisé d'obtenir une transcription du corpus en français avec Whisper un modèle de transcription automatique développé par OpenAI. On dispose

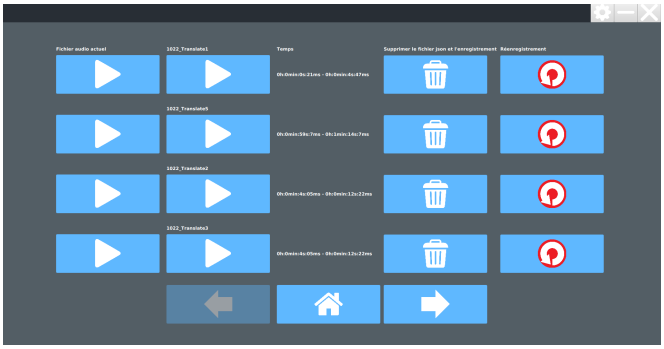


FIGURE 2 – L'écran de validation des interprétation de TètKole

alors de la parole en haïtien, de la parole en français et de la transcription en français.

4 Perspectives Futures

4.1 Valorisation des enregistrements de terrain

Des années de travail de terrain par une multitude de linguistes ont permis de collecter un grand nombre d'enregistrements, dont seule une petite fraction a été transcrite. Ces enregistrements n'ont généralement pas été valorisés par le passé, faute d'une solution technique appropriée. L'avénement des modèles d'apprentissage auto-supervisé (*self-supervised learning*, SSL) qui ne nécessitent pas de transcriptions permettent désormais d'envisager une valorisation de ces données.

Nous souhaitons donc étudier l'utilisabilité des données de terrain *déjà existantes* dans cadre d'un pré-entraînement auto-supervisé de modèles WAV2VEC2. L'utilisation de ces données représente un défi, et ce à deux titres. D'une part, ces données ont généralement été enregistrées sur cassettes, puis digitalisées, et sont donc de qualité acoustique variable. D'autre part, ces enregistrements font suite à des enquêtes de terrain répondant à des problèmatiques scientifiques diverses (par exemple des enquêtes grammaticales, atlas sonore, etc.) et ne rentrent pas dans le canon standard des données actuellement utilisées pour entraîner de tel modèles (généralement des livres audios) contenant en grande majorité de parole lue.

Ainsi nous souhaitons explorer l'influence de plusieurs variables sur la qualité des modèles entraînés. Plus précisément, nous explorerons l'influence de la nature des données de terrain, de la qualité acoustique, du ratio parole lue/parole spontanée, et de la quantité de données minimale nécessaire. Ces questions restent à ce jour encore ouvertes, surtout dans le cadre des langues peu dotées.

4.2 Des modèles neuronaux et des questionnements linguistiques

Les modèles neuronaux que nous entraînerons seront mis au service d'un questionnement linguistique. Ainsi, ces modèles nous permettront d'explorer l'influence des langues lexificatrices [3] sur les créoles. Pour explorer cette question, nous souhaitons donc entraîner des modèles auto-supervisés "pan-créoles" (incluant par exemple des données d'haïtien, de martiniquais, de réunionais, etc.) à partir de rien (*from scratch*), à partir de leur langue lexificatrice, à partir d'une autre langue, ou bien à partir de plusieurs langues (en utilisant des modèles multilingues, de type XLSR (Conneau *et al.*, 2021)). Cela nous permettra ainsi de voir si les créoles partagent plus de traits entre-eux qu'avec leur langue lexificatrice, auquel cas le modèle pan-créole se révélerait meilleur une fois raffiné (*fine-tuned*) sur chaque créole pris indépendamment, ou non. Les résultats de Macaire *et al.* (2022) semblent indiquer qu'il est souhaitable d'entraîner des modèles de reconnaissance de parole à partir de la langue lexificatrice plutôt que de modèles multilingues. Cependant, cela reste à vérifier sur des créoles ayant une autre langue lexificatrice.

Références

AUSTIN P. K. (2016). Language documentation 20 years on. *Endangered languages and languages in danger : Issues of documentation, policy, and language rights*, p. 147–170.

BAEVSKI A., ZHOU Y., MOHAMED A. & AULI M. (2020). wav2vec 2.0 : A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, **33**, 12449–12460.

BLACHON D., GAUTHIER E., BESACIER L., KOUARATA G.-N., ADDA-DECKER M. & RIALLAND A. (2016). Parallel speech collection for under-resourced language studies using the lig-aikuma mobile device app. *Procedia Computer Science*, **81**, 61–66.

CHAUDENSON R. (2005). Description et graphisation : le cas des créoles français. *Revue française de linguistique appliquée*, **10**(1), 91–102.

CONNEAU A., BAEVSKI A., COLLOBERT R., MOHAMED A. & AULI M. (2021). Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, p. 2426–2430. DOI : 10.21437/Interspeech.2021-329.

EVAIN S., NGUYEN H., LE H., BOITO M. Z., MDHAFFAR S., ALISAMIR S., TONG Z., TOMASHENKO N., DINARELLI M., PARCOLLET T. *et al.* (2021). Lebenchmark : A reproducible framework for assessing self-supervised representation learning from speech. In *INTERSPEECH 2021 : Conference of the International Speech Communication Association*.

3. On désigne par *langue lexificatrice*, les langues "dont les créoles ont retenu la plus grande partie de leur lexique si ce n'est de leur grammaire" (Mufwene, 2021)

GRAVES A., FERNÁNDEZ S., GOMEZ F. & SCHMIDHUBER J. (2006). Connectionist temporal classification : Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, p. 369–376, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/1143844.1143891.

LEHMANN C. (2001). *Language documentation : A program.* na.

MACAIRE C., SCHWAB D., LECOUTEUX B. & SCHANG E. (2022). Automatic speech recognition and query by example for creole languages documentation. In *Findings of the Association for Computational Linguistics : ACL 2022*, p. 2512–2520.

MICHAUD A., GUILLAUME S., JACQUES G., MAC D.-K., JACOBSON M., PHAM T.-H. & DEO M. (2016). Contribuer au progrès solidaire des recherches et de la documentation : la collection pangloss et la collection auco. In *Journées d'Etude de la Parole 2016*, volume 1, p. 155–163.

MUFWENE S. (2021). Créoles. *Langage et société*, **Hors série**(HS1), 81–86. DOI : 10.3917/ls.hs01.0082.

Actes des journées LIFT 2023
Nancy, France, 20-21 novembre 2023

Plus que des données: La question de la variation dialectale et les ressources en TAL pour le corse et le poitevin-saintongeais

Cristina Garcia Holgado[1,2]
(1) FoReLLIS, UMR 15076, Université de Potiers, 86000 Poitiers, France
(2) LISA, UMR CNRS 6240, Université de Corse - Pasquale Paoli, Corte, France
cristina.garcia.holgado@univ-poitiers.fr

RÉSUMÉ

Le poitevin-saintongeais et le corse ont récemment rejoint la communauté du TALN. Cependant, ces langues sont confrontées à des défis importants en raison de la rareté des corpus annotés, et en outre, elles ne constituent pas des entités uniformes mais présentent des variétés dialectales multiples, malgré les efforts controversés pour établir une référence standardisée. Au cours de la dernière année, elles ont été dotées de ressources linguistiques qui ont conduit aux premières tentatives d'exploration et d'évaluation d'un nombre d'applications TALN. Cependant, mettre en lumière une langue régionale dans le paysage technologique implique également de reconnaître sa réalité linguistique : la diversité dialectale à travers ses différents territoires. Dans ce résumé, nous présentons les premières expériences d'application de méthodes et de techniques supervisées pour gérer le manque d'annotations, en particulier pour le poitevin-saintongeais, et nous soulignons l'importance de couvrir leur variation dialectale, ce que nous cherchons à aborder dans nos travaux futurs.

ABSTRACT

More than just data : Dialectal variation and NLP resources for Corsican and Poitevin-Saintongeais

Poitevin-Saintongeais and Corsican have recently joined the NLP community. However, they face significant challenges due to the scarcity of annotated corpora, and moreover, they do not constitute a homogeneous entity but multiple dialectal varieties, despite the controversial efforts to establish a standardized reference. In the past year, they have been equipped linguistic resources, leading to the first attemps to explore and evaluate a few NLP methods. However, shedding light on a regional language in the technological landscape also entails acknowledging their linguistic reality : the dialectal diversity across it's differents territories. In this summary, we outline initial experiences applying supervised methods and techniques to handle the lack of annotations, specially for Poitevin-Saintongeais, and we highlight the importance of covering their dialectal variation, which we seek to address in future work.

MOTS-CLÉS : corse, poitevin-saintongeais, langues régionales, langues peu dotées, lexiques, variation dialectale.

KEYWORDS: corsican, poitevin-saintongeais, regional languages, low resource language, lexicons, dialectal variation.

1 Context

There is a growing interest in providing digitalised linguistic resources to regional languages in France as shown in (Kevers *et al.*, 2019), (Millour *et al.*, 2017) and (Bernhard *et al.*, 2021). In the case of Corsican and Poitevin-Saintongeais, numerous textual resources are available : The first benefits from an online linguistic database, the BDLC (Banque de Données de la Langue Corse) (Stella Retali-Medori, 2022) which originally included texts from oral sources from different regions, and currently integrates the CCdC (Corpus Canopé de Corse, (Kevers, 2022)) composed of literary and historical texts. Poitevin-Saintongeais benefits from the TELPOS (Dourdet *et al.*, 2019) database, which contains more than 125 bibliographic references for literary texts. The textual resources for Poitevin-Saintongeais are characterized by different spellings, where only a few use the standard spelling. Moreover, a few aligned fragments with other regional languages are available in the ParCoLaF (Miletic *et al.*, 2017) database [1] for both languages. Besides the avaliability of these digital resources in both languages, most of that data remains unannotated and not readily exploitable. In the last year, these two languages have ventured into the NLP domain seeking to develop annotated corpus and lexicons to experiment with supervised approaches.

2 Recent work

2.1 Difficulties in a low resource scenario

Recent approaches in low resource (LR) settings generally rely on a high resource related language, using methods like transfer learning by choosing a suitable transfer language. In this sense, while it may seem intuitive to utilize French for Poitevin-Saintongeais and Italian for Corsican, given their shared Gallo-Romance and Italo-Romance roots respectively, these languages encounter specific challenges to be considered :

1. A **limited availability of linguistic resources**, such as lexicons and dictionaries. Although a few exist, they are subject to copyright whose access is not always guaranteed, or they are based on a particular spelling.
2. **Limited parallel corpus**. Aligned sentences from two literary works are available at ParCoLab, but the amount of aligned data remains very scarce for both languages. However, Poitevin-Saintongeais profits from bilingual articles featured in the journal

1. http ://parcolab.univ-tlse2.fr/corpus/search

Culture Nouvelle Aquitaine [2] since 2023, with uses standardized spelling and provides an opportunity to increase and exploit parallel corpora.

3. Generally, regional languages **lack of standardized spelling** which adds a significant complexity, specially for tasks that are necessary in the area of descriptive linguistics, such as lemmatization. Both Corsican and Poitevin-Saintongeais have a normalized spelling, and the first has been recognized as co-official language in Corsica since 2013. However, the texts available in these languages are very diverse, belonging to different geographical areas and therefore, characterised by different variants, speeches (*parlers*) and spellings (*graphie*) that must be taken into account.
4. Along with this, an important challenge arises from the **diverse diffusion areas** (*aires dialectales*) found in regional languages. When we refer to data scarcity, we also encompasses the scarcity of available texts that are annotated with consideration to their various dialectal features that actually constitute those languages. To date, the texts annotated for Poitevin-Saintongeais have been limited to those with standardized spelling so as to address the dialectal variety question in a later stage. For corsican, a set of texts from different sources were annotated regardless of the presence of dialectal variation.
5. Although the BDLC, and particularly the TELPOS database, contain an important number of texts, there is **insufficient metadata** for an effective dialect characterization and processing from an NLP perspective. While a few metadata information on Poitevin-Saintongeais, such as the use of standardized spelling or locality, has been sporadically annotated in some texts, the same attributes are currently unavailable for Corsican texts. Hence, there is still considerable work to annotate these texts in the databases to provide them with linguistic and geographical metadata.

2.2 Approaches

Early work on these languages has focused on overcoming the lack of annotated data. An annotation campaign took place during the last year leading to the first NLP experiences for both languages, which primarily targeted transfer learning methods considering the available resources :

— For Corsican, morpho-syntactic analysis started to be evaluated at LISA (University of Corsica) over an annotated corpus of ~7k tokens. This work, which is in progress, assested the effectiveness of pretraining embeddings in corsican and italian, while using different training sizes. Additional resources, such as parallel corpuses or lexicons, were not used at present, but it could be an envisageable option for the future considering the progressive availability of parallel sentences in ParCoLaF and the constitution of a morphologically inflected lexicon.

2. https ://www.culture-nouvelle-aquitaine.fr/langues-et-cultures-regionales/traduire-le-site-en-langues-regionales/

— Poitevin-saintongeais has followed a similar path including lemmatization. A smaller corpus of ~3k tokens was used, which made this task more arduous. To address this data gap, a lexicon of ~20k entries was compiled via an online bilingual dictionary (Pivetea, 2006), transformed, expanded (~40k entries) and adapted to the respective Universal Dependencies (UD) guidelines. This work had a dual goal : first, to accelerate annotator's decisions by integrating the lexical entries to a collaborative annotation notebook, and second, to generate augmented corpora by transferring new lexical information via distributional neighbours to assess the benefits of a lexicon-based strategy for morpho-syntactic analysis, using both probabilistic (HMM) and neural models (LSTM). This approach has proved to be beneficial without requiring an extensive lexicon. An improvement will be expected with the incorporation of inflected verbs, which were not naturally present in the source dictionary as opposed to nouns and adjectives for which we could provide the inflectional paradigm.

2.3 Limitations

Both lines of work sought to increase the number of annotated texts in order to be able to perform finer NLP tasks, but also to propose the first pos-tagging models for these languages. However, several questions arise at this point : When enlarging the annotated corpus, how well are the distinct linguistic phenomena of these languages represented, and how does the representation of their syntactic and morphological structures impact the quality of predictions ? And most important, how effective are these models when applied to their different dialectal variants ? These questions are intended to show that the quantity of annotated texts is not a sufficient objective in the framework of regional languages, but also the quality.

3 Conclusion and future work

While basic NLP tasks seem straightforward due to the availability of different methods that gradually try to adapt to low resource scenarios, a major challenge arises when addressing dialectal variation. Despite the positive results of this work, none of them have taken yet into account their dialectal dimension. Given the nature of the project they integrate, this stage therefore becomes a required line of research in the work to come. This would enable the representation of their linguistic reality by offering a more nuanced visibility into their different geographical areas. These differences are primarily evident in the lexical and phonetic levels, although they can extend to the morphology and to the idiomatic expressions. In this context, we consider essential to understand that the efforts dedicated to equip these languages go beyond providing enough data for NLP applications. They are motivated by the broader goal of preserving and revitalizing their linguistic heritage. As a result, this undertaking necessitates a comprehensive understanding of their intricate linguistic realities,

and for that, the NLP community requires a strong support from linguistic experts to cover this essential feature. In this sense, the first objective will be to characterize and to identify the variation, and to do so, we will require a significant effort in representing these varieties in the corpus. In short, handling linguistic variation is a central focus of our current work, which has been embedded within a thesis project that will seek to develop tools to ensure their survival and growth, as both Corsican and Poitevin are considered endangered languages by the UNESCO.

4 Acknowledgements

This work was funded by the National Research Agency, via the project DIVITAL (ANR-21-CE27-0004).

Références

BERNHARD D., LIGOZAT A.-L., BRAS M., MARTIN F., VERGEZ-COURET M., ERHART P., SIBILLE J., TODIRASCU A., DE MAREÜIL P. B. & HUCK D. (2021). Collecting and annotating corpora for three under-resourced languages of france : Methodological issues. *Language Documentation & Conservation*, **15**, 316–357.

DOURDET J.-C., VERGEZ-COURET M. & LAY M.-H. (2019). Telpos - Texte électronique en poitevin-saintongeais, enjeux et difficultés. In *Colloque "Langues minoritaires" : quels acteurs pour quel avenir ?*, Strasbourg, France. HAL : hal-02892750.

KEVERS L. (2022). *CCdC - Le Corpus Canopé de Corse*. Rapport interne, UMR 6240 CNRS LISA - Université de Corse. HAL : hal-03912288.

KEVERS L., GUÉNIOT F., GHJACUMINA TOGNOTTI A. & RETALI MEDORI S. (2019). Outiller une langue peu dotée grâce au TALN : l'exemple du corse et BDLC. In E. MORIN, S. ROSSET & P. ZWEIGENBAUM, Éds., *26e Conférence sur le Traitement Automatique des Langues Naturelles*, p. 371–380, Toulouse, France : ATALA. HAL : hal-02567779.

MILETIC A., STOSIC D. & MARJANOVIĆ S. (2017). Parcolab : A parallel corpus for serbian, french and english. In *International Conference on Text, Speech and Dialogue*.

MILLOUR A., FORT K., BERNHARD D. & STEIBLÉ L. (2017). Vers une solution légère de production de données pour le TAL : création d'un tagger de l'alsacien par crowdsourcing bénévole. In *Traitement Automatique des Langues Naturelles (TALN)*, Orléans, France. HAL : hal-01516226.

PIVETEA V. (2006). *Dictionnaire francais poitevin-saintongeais*. Geste Éditions.

STELLA RETALI-MEDORI L. K. (2022). La morphologie dans la banque de données langue corse : bilan et perspectives. *OpenEdition*.

Actes des journées LIFT 2023
Nancy, France, 20-21 novembre 2023

Prise en compte de la variation dans l'annotation automatique morphosyntaxique de l'occitan

Clamença Poujade[1, 2]
(1) CLLE, UMR 5263, Université de Toulouse Jean Jaurès, CNRS
5, allées Antonio Machado, 31078 cedex 9 Toulouse, France
(2) JOLICIEL Informatique, 2 Av. du Cardie, 09000 Foix, France
clamenca.poujade@univ-tlse2.fr

RÉSUMÉ

L'occitan est une langue romane de France, d'une petite partie de l'Italie et de l'Espagne. Il comprend de nombreuses variations à l'écrit, notamment les variations dialectale et de graphie. C'est un enjeu important dans la dotation de la langue que de pouvoir prendre en compte la variation. Le traitement automatique de l'occitan est en développement cette dernière dizaine d'années. Des ressources et des outils sont constitués et commencent à prendre en compte la variation dialectale. Toutefois, la variation graphique est peu présente dans ces travaux. Notre travail de recherche se concentre sur l'annotation automatique en lemmes, en parties du discours et en flexion verbale d'un corpus de textes contenant ces deux types de variation. À partir de ce corpus nous entraînons des outils d'annotation automatique robustes sur la variation globale de l'occitan.

ABSTRACT

Variation in Automatic Annotation of Occitan.

Occitan is a Romance language of France, a little part of Italy and Spain. It includes many written variations, dialectal and spelling variations. Being able to take variation into account is a major challenge to provide the language. Automatic processing of Occitan has been developing over the last ten years. Resources and tools have been developed and are beginning to take dialectal variation into account in these works. However, graphical variation is rarely taken into account. Our research focuses on the automatic annotation into lemmas, parts of speech and verbal inflection of a corpus of texts containing these two types of variation. From this corpus we train robust automatic annotation tools on global variation in Occitan.

MOTS-CLÉS : Annotation automatique - Variation - Langue moins dotée - occitan - parties du discours - corpus.

KEYWORDS: Automatic Annotation - Variation - Less-ressourced Language - Occitan - parts of speech - corpus.

FIGURE 1 – Carte des dialectes de l'occitan (Bernhard *et al.*, 2021)

L'occitan est une langue parlée dans le sud de la France, dans une vallée des Pyrénées en Espagne et dans certaines vallées du Piémont en Italie. Il compte six grands dialectes (Bec, 1995) (Figure 1) et ces dialectes sont constitués de nombreux parlers. La distinction entre les dialectes se fait via des faisceaux d'isoglosses créant un continuum linguistique.

1 Ressources et outils pour la linguistique outillée de l'occitan, état des lieux

L'occitan a longtemps été considéré comme une langue peu dotée dans le traitement automatique des langues. Toutefois, depuis une dizaine d'années, la langue occitane s'est dotée de plusieurs corpus numériques et de corpus annotés, d'outils numériques et automatiques de traitement de la langue. Notre groupe de recherche sur l'outillage des langues peu dotées [1] travaille à sa dotation, et un organisme associatif de régulation de la langue produit des applications pour le grand public [2].

Parmi ces ressources, l'occitan dispose d'une base textuelle, BaTelÒc (Bras & Vergez-Couret, 2016) constituant un corpus de presque quatre millions de mots, mis en ligne et interrogeable via une interface. Notre groupe de recherche a également construit un corpus annoté en Parties du discours (POS) dans le cadre du projet ANR Restaure (Bernhard *et al.*, 2018) en collaboration avec d'autres langues de France et un corpus annoté en POS et en dépendances syntaxiques (Tolosa Treebank) dans le cadre du projet européen Linguatec, avec d'autres langues des Pyrénées (Miletic *et al.*, 2020a). En comparaison avec des langues bien dotées, ces corpus sont de petite taille, ils comptent seulement quelques dizaines de milliers de mots. Ces deux corpus annotés ont pris en compte une certaine variation linguistique avec la présence de quatre dialectes dans le corpus Restaure et de cinq dialectes dans le corpus Linguatec (Miletic *et al.*, 2020b). En collaboration avec Lo Congrès, nous avons également construit un lexique de formes fléchies (Bras *et al.*, 2020) qui a servi à l'entraînement de premiers outils d'annotation automatique pour l'occitan. Les outils

1. Groupe OCRE : https ://clle.univ-tlse2.fr/accueil/equipes-de-recherche/sciences-du-langage/occitan-langues-romanes-langues-deurope-decrire-formaliser-outiller-comparer
2. Lo Congrès Permanent de la lenga occitana : https ://locongres.org/

entraînés avec ces ressources fournissent de bons résultats pour l'annotation en parties du discours et dépendances syntaxiques des différents dialectes ; mais ces ressources ne contiennent pas de variation graphique.

2 Enjeux du traitement de la variation

L'occitan est une langue n'ayant pas de standard vraiment défini. Beaucoup de textes sont écrits dans un parler qui est propre à l'auteur ou l'autrice.

La graphie non plus n'a pas de norme générale arrêtée. Cependant, plusieurs normes se font concurrence : la graphie "classique" (Exemple 1), qui est celle construite avec l'objectif d'avoir une graphie propre à la langue (et qui est celle présente dans les corpus annotés occitans déjà construits) ; la graphie dite "mistralienne" (Exemple 2) qui est, en partie, construite à partir de la norme orthographie française ; et les graphies personnelles (Exemple 3) des auteur·rice·s qui sont, souvent, des graphies dites oralisantes.

'Tous les chiens sont heureux.'

1. Totes los gosses son uroses.
2. Toutés lous goussés soun urousés.
3. Toutéÿ louÿ goussés soun urouzés.

Ces variations rendent difficile l'exploitation de tous les textes occitans. Afin de pouvoir les exploiter et en tirer des informations linguistiques, il est nécessaire de construire des outils de traitement automatique qui soient robustes pour traiter ces variations. Nous pensons que les nouvelles générations d'outils automatiques, par exemple les architectures utilisant des réseaux de neurones, se montrent plus robustes sur la variation que les anciennes générations, comme les modèles d'apprentissage supervisé par règles.

3 Nouveau corpus de textes occitans pour travailler sur la variation

Ce travail s'inscrit dans cette nécessité d'entraîner des outils pour qu'ils prennent en compte la variation afin qu'ils soient robustes sur l'ensemble des données. Nous avons choisi de nous concentrer sur la construction d'un corpus de textes du département de l'Ariège.

3.1 La variation dialectale

Ce département se trouve à la frontière sud de la France avec l'Espagne. C'est une zone de transition linguistique où l'on trouve de nombreuses isoglosses entre deux dialectes occitans,

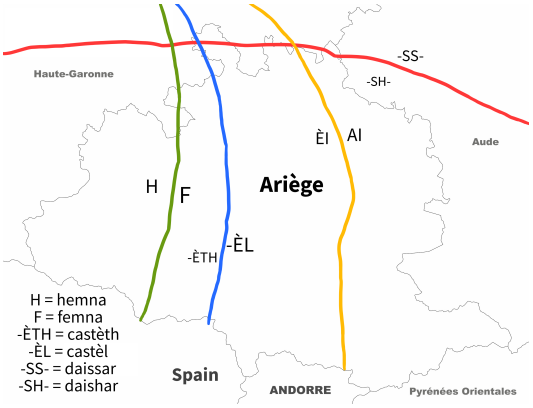


FIGURE 2 – Carte de quelques isoglosses passant par l'Ariège

le gascon (à l'ouest) et le languedocien (à l'est) (Figure 2 [3]). Dans ce département nous trouvons également des parlers "isolés", dans les montagnes, qui ne font pas partie de la zone de transition entre les deux dialectes, mais plutôt entre deux langues : l'occitan et le catalan. Cela fait de ce petit territoire occitan une zone présentant une grande variation dialectale.

Ce n'est pas la seule variation qui nous intéresse dans cette recherche, l'Ariège a aussi la chance de disposer d'une grande production d'écrits occitans. De nombreux·se·s auteur·rice·s ont écrit des contes, légendes, romans ou même des articles de journal en occitan, dans leur parler ariégeois.

3.2 La variation graphique

Comme nous l'évoquions plus haut, l'occitan a plusieurs graphies. Dans ce département nous ne pouvons pas quantifier le nombre de graphies existantes, mais nous pouvons les regrouper dans les trois catégories décrites plus haut.

Les locuteur·rice·s de l'occitan n'étant que très rarement alphabétisé·e·s en occitan, nous trouvons de nombreuses personnes ayant l'envie d'écrire dans leur langue mais ne connaissant pas les normes qui peuvent être utilisées. Elles font alors avec les connaissances qu'elles ont, en l'occurrence, en utilisant les graphèmes de la langue dominante, ici, le français. Ces graphies oralisantes sont très intéressantes à étudier afin de mieux connaître la phonologie ou la morphologie de leurs parlers à partir d'écrits.

3.3 Constitution du corpus

Afin de construire de tels outils, nous avons constitué un corpus de textes pertinents. Nous avons rassemblé des textes issus des différents groupes de parlers de l'Ariège et des trois

3. Réalisée à partir du fond de carte https ://d-maps.com/carte.php ?num_car=111145&lang=fr

# tokens	Graphie		
Dialecte	Classique	Mistralienne	Toutes
Languedocien	10 644	7 112	17 756
Gascon	5 710	6 749	12 459
Autre	2 936	5 163	8 099
Tout	19 290	19 024	38 314

FIGURE 3 – Répartition des tokens dans le corpus Arièja

```
Un      un      un      DET     Gender=Masc|Number=Sing
joun    joun    jorn    NOUN    Gender=Masc|Number=Sing
me      me      me      PRON    Number=Sing|Person=1
passèc  passa   passar  VERB    Number=Sing|Person=3|Mood=Ind|Tense=Past
uno     un      un      DET     Gender=Fem|Number=Sing
ideio   ideio   idèia   NOUN    Gender=Fem|Number=Sing
pel     _       _       _       _       _
per     per     per     ADP     _       _
le      le      lo      DET     Gender=Masc|Number=Sing
cap     cap     cap     NOUN    Gender=Masc|Number=Sing
.       .       .       PUNCT   _       _
```

FIGURE 4 – Annotation d'une phrase du corpus Arièja

types de graphies. Nous avons choisi de regrouper la graphie mistralienne et les graphies oralisantes étant donné qu'elles sont basées sur le rapport graphie-phonie du français.

Notre corpus (Corpus Arièja) est représentatif de toute ces variations théoriquement présente sur le territoire. Certains textes sont assez récents pour en avoir une copie numérique, mais ce n'est pas le cas de tous. Avant de pouvoir les traiter, nous avons dû les numériser dans la quasi entièreté et les océriser.

Le Corpus Arièja est constitué de 38 314 tokens avec 66 textes et 47 auteurs et autrices. Le tableau 3 présente la répartition des tokens dans le corpus en fonction des dialectes et des graphies des textes. Nous avons gardé des proportions similaires pour les deux graphies mais il n'a pas été possible de faire la même chose pour les dialectes.

4 Annotation du corpus et outils

Le corpus est annoté en lemmes, parties du discours (POS) et la flexion verbale l'est également. Pour les annotations POS et de la flexion verbale, nous suivons les préconisations de Universal Dependencies (Nivre *et al.*, 2016).

L'annotation des lemmes se fait en deux parties. Un premier lemme qui suit la graphie de l'auteur·rice : nous déduisons du reste du texte le lemme que l'auteur·rice aurait produit. Nous annotons également ce que nous appelons le Supra-lemme, qui est un lemme qui ne suit pas toujours la graphie de l'auteur·rice. Ce Supra-lemme sert simplement à accéder plus facilement aux tokens qui nous intéressent sans que la variation n'interfère. L'illustration 4 montre les différentes annotations présentes dans le corpus.

Restaure - Tolosa Treebank

a) Talismane

Dialecte	% Exactitude
Languedocien	89,49
Limousin	82,48
Gascon	81,41
Provençal	81,37
Tout	83,32

b) AllenÒc

Dialecte	% Exactitude
Languedocien	97,30
Provençal	96,72
Limousin	94,47
Gascon	93,65
Tout	96,75

Corpus Arièja

a) Talismane

% Exactitude	Graphie		
Dialecte	Classique	Mistralienne	Toutes
Languedocien	82,37	67,27	76,09
Gascon	76,25	67,82	72,50
Autre	/	68,6	68,6
Tout	79,65	67,98	73,23

b) AllenÒc

% Exactitude	Graphie		
Dialecte	Classique	Mistralienne	Toutes
Languedocien	94,44	89,79	92,74
Gascon	90,60	82,89	87,37
Autre	/	90,71	90,71
Tout	93,26	91,32	93,35

FIGURE 5 – Résultats des modèles sur différents corpus

Le corpus est annoté, pour partie (21 301 tokens) manuellement (à partir d'une pré annotation automatique) et pour partie (17 013 tokens) automatiquement, avec les outils finalisés et ayant de bons résultats sur les variations du corpus.

Nous comparons les résultats de deux modèles d'apprentissage automatique pour déterminer celui qui est le plus robuste face à cette variation. L'un, Talismane (Urieli, 2013; Vergez-Couret & Urieli, 2015) est une architecture qui avait déjà servi pour l'annotation des corpus issus des projets Restaure et Linguatec. C'est un modèle d'apprentissage par règles et probabiliste. L'autre, AllenÒc, utilise des réseaux de neurones et est basé sur la bibliothèque de deep-learning AllenNLP (Gardner *et al.*, 2018).

Ces premiers modèles d'apprentissage automatique, dont nous donnons les résultats ci-après, sont entraînés à partir des corpus Tolosa Treebank (Miletic *et al.*, 2020a) et Restaure (Bernhard *et al.*, 2018) n'intégrant pas de variation graphique, pour faire de l'annotation automatique en parties du discours. Ils seront une nouvelle fois entraînés avec la partie du CorpusArièja corrigée manuellement et qui inclut plusieurs graphies. Ces modèles ont été entraînés à partir de peu de données. Toutefois plusieurs études (Bernhard *et al.*, 2021), ainsi que nos résultats, montrent que nous pouvons obtenir de bons scores d'annotation automatique avec peu de données d'entraînement. Nous montrons que, même avec peu de données d'entraînement, nous pouvons avoir des outils robustes et efficaces pour l'annotation de la plupart des données textuelles.

Les résultats (Figure 5) montrent que le modèle AllenÒc est bien plus performant que le modèle Talismane, que ce soit sur la variation dialectale ou graphique. Les deux modèles ont tendance a être plus performants sur le languedocien et moins sur le gascon. Cela s'explique par le nombre de tokens différents de chacun des dialectes dans les corpus d'entraînement. Le dialecte le plus présent est le languedocien et les autres dialectes sont moins présents. Le test sur le corpus Arièja, avec les différentes graphies, montre que le modèle AllenÒc est plus robuste face aux variations graphiques que Talismane.

Une fois les modèles entraînés sur des corpus contenant de la variation graphique, nous nous attendons à avoir de meilleurs résultats sur la variation graphique. Pour ce qui est d'AllenÒc, nous pensons que les résultats seront bien meilleurs, et rejoindront presque ceux que nous avons sur les textes sans variation graphique. Toutefois, en ce qui concerne Talismane, nous pensons que les résultats ne seront pas beaucoup améliorés. En effet, il est très sensible à la variation au sein de son corpus d'entraînement.

Références

BEC P. (1995). *La langue occitane*. Que sais-je ? 1059. Paris : Presses universitaires de France, 6e édition corrigée.

BERNHARD D., LIGOZAT A.-L., BRAS M., MARTIN F., VERGEZ-COURET M., ERHART P., SIBILLE J., TODIRASCU A., BOULA DE MAREÜIL P. & HUCK D. (2021). Collecting and annotating corpora for three under-resourced languages of France : Methodological issues. *Language Documentation & Conservation*, **15**, 316–357. HAL : hal-03273196.

BERNHARD D., LIGOZAT A.-L., MARTIN F., BRAS M., MAGISTRY P., VERGEZ-COURET M., STEIBLE L., ERHART P., HATHOUT N., HUCK D., REY C., REYNÉS P., ROSSET S., SIBILLE J. & LAVERGNE T. (2018). Corpora with Part-of-Speech Annotations for Three Regional Languages of France : Alsatian, Occitan and Picard. In *11th edition of the Language Resources and Evaluation Conference*, Miyazaki, Japan. HAL : hal-01704806.

BRAS M., HATHOUT N., SIBILLE J., VERGEZ-COURET M., SÉGUIER A. & DAZEAS B. (2020). Loflòc : Lexic Obert flechit occitan. In J.-F. C. ET DAVID FABIÉ, Éd., *Fidelitats e dissidéncias. Actes del XIIn Congrès de l'Associacion internacionala d'estudis occitans. Actes du XIIe Congrès de l'Association internationales d'études occitanes. Albi 10-15/07/2017*, p. 141–156. Section française de l'Association internationale d'Etudes Occitanes. HAL : hal-03082686.

BRAS M. & VERGEZ-COURET M. (2016). Batelòc : A text base for the occitan language.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

GARDNER M., GRUS J., NEUMANN M., TAFJORD O., DASIGI P., LIU N., PETERS M., SCHMITZ M. & ZETTLEMOYER L. (2018). Allennlp : A deep semantic natural language processing platform.

MILETIC A., BRAS M., VERGEZ-COURET M., ESHER L., POUJADE C. & SIBILLE J. (2020a). Building a Universal Dependencies Treebank for Occitan. In *12th Language Resources and Evaluation Conference*, p. 2932–2939, Marseille, France. HAL : hal-02892715.

MILETIC A., BRAS M., VERGEZ-COURET M., ESHER L., POUJADE C. & SIBILLE J. (2020b). A four-dialect treebank for Occitan : Building process and parsing experiments. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, p. 140–149, Barcelona, Spain (Online) : International Committee on Computational Linguistics (ICCL).

NIVRE J., DE MARNEFFE M.-C., GINTER F., GOLDBERG Y., HAJIČ J., MANNING C. D., MCDONALD R., PETROV S., PYYSALO S., SILVEIRA N., TSARFATY R. & ZEMAN D. (2016). Universal Dependencies v1 : A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 1659–1666, Portorož, Slovenia : European Language Resources Association (ELRA).

URIELI A. (2013). *Robust French syntax analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Theses, Université Toulouse le Mirail - Toulouse II. HAL : tel-00979681.

VERGEZ-COURET M. & URIELI A. (2015). Analyse morphosyntaxique de l'occitan languedocien : l'amitié entre un petit languedocien et un gros catalan. In *TALARE 2015*, Caen, France. HAL : hal-01214566.

Actes des journées LIFT 2023
Nancy, France, 20-21 novembre 2023

Segmentation et analyse de productions non modales, une étude de cas : la langue korebaju

Jenifer Vega Rodriguez[1,2] Nathalie Vallée[1]
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France
(2) Université de Brasilia, Brasilia, Brésil.
Jenifer-andrea.vega-rodriguez@gipsa-lab.fr, nathalie.vallée@gipsa-lab.fr

RESUME

Le korebaju (koreguaje) est une langue tonale appartenant à la famille Tukano parlée dans l'Amazonie colombienne. Cette langue présente plusieurs types de productions non modales impliquant une constriction glottale : [ʔ̰, V̰, Vˀ, ˀV, Vˀ, ˀV, ʔ] et formant des séquences telles que (C)VʔV, (C)V̰CV, (C)Vʔ̰CV ou (C)VʔCV, indépendamment de la hauteur tonale. De plus, ces productions présentent des variations de réalisation intra- et interlocuteur.

Ce document présente l'ensemble des productions glottales relevées dans deux variétés du korebaju (korebaju et tama), sur la base de données audio et électroglottographiques, et recueillies sur le terrain entre 2022 et 2023 auprès de locuteurs natifs. Nous souhaitons présenter l'inventaire et la distribution des différents types de productions glottales en proposant leurs caractérisations acoustiques compte tenu des variations intra- et interlocuteurs pour ainsi ouvrir une discussion avec des experts en traitement automatique de la parole en vue du développement d'un outil de segmentation et labélisation (semi) automatique des corpus korebaju.

ABSTRACT

Segmentation and analysis of non-modal productions, a case study: the Korebaju language. Korebaju (Koreguaje) is a tonal language that belongs to the Tukano family spoken in the Colombian Amazon. This language presents several types of non-modal production involving glottal constriction [ʔ̰, V̰, Vˀ, ˀV, Vˀ, ˀV, ʔ] and occurring in sequences like (C)VʔV, (C)V̰CV, (C)Vʔ̰CV or (C)VʔCV regardless of tone height. In addition, they show intra- and inter-speaker variability.

Our paper presents these glottal productions identified in two Korebaju varieties (Korebaju and Tama) on the basis of audio and electroglottographic data collected in the field between 2022 and 2023 from native speakers. Our aim is to present the inventory and distribution of the different types of glottal productions, and to propose their acoustic characterizations, taking into account intra- and inter-speaker variabilities, in order to open a discussion with experts in automatic speech processing with a view to developing a tool for (semi-) automatic segmentation and labeling of Korebaju corpora.

MOTS-CLES : Glottalisation, phonétique et phonologie, langues tukano, koreguaje.
KEYWORDS: Glottalization, phonetics and phonology, Tukanoan Languages, Koreguaje.

1 Introduction

Le korebaju [kòˀréβ̞àhɨ́], appelé également koreguaje, est une langue tonale de la famille Tukano (branche occidentale) parlée dans le sud-ouest de l'Amazonie colombienne. À ce jour, la communauté Korebaju est constituée d'environ 2 000 membres disséminés dans 27 villages (Communauté Korebaju, 2011). Historiquement parlant, elle s'est constituée avec le regroupement de quatre communautés : Korebaju, Tama, Macaguaje et Carijona, chacune ayant sa propre langue. À partir du XVI^{e} siècle, les processus d'esclavagisme et d'évangélisation vont progressivement modifier cette configuration et amener les quatre clans à adopter la langue korebaju comme leur propre langue. Actuellement, la communauté Korebaju présente des variations dialectales souvent bien identifiées par les locuteurs eux-mêmes (nous citons en exemple le propos du cuisinier tama qui nous avertit lors de l'enquête qu'il ne parle que le tama). L'existence d'une variation dialectale est également mentionnée par quelques études descriptives précédentes (p. ex. Cook & Criswell, 1993) sans fournir toutefois beaucoup de précisions, ni consensus sur sa nature. Le korebaju est aujourd'hui menacé d'extinction et répertorié dans l'« Atlas des langues en danger dans le monde » de l'UNESCO (Moseley, 2010). Nous avons conduit une étude récente sur ces données d'enquête qui nous a permis de proposer 15 consonnes /p, t, k, p^{h}, t^{h}, k^{h}, β, ɸ, s, h, w, t͡ʃ, m, n, ɲ, ^{h}ɲ, r/ et 17 voyelles /i, e, a, o, u, ɨ, ĩ, ẽ, ã, õ, ũ, ɨ̃, i^{ʔ}, e^{ʔ}, a^{ʔ}, o^{ʔ}, ɨ^{ʔ}/ pour le korebaju (Vega Rodriguez et al., 2023) confirmant ainsi certains des résultats d'une étude prélimaire de Vega Rodriguez (2019).

Dans la problématique de la description phonétique et phonologique du korebaju, une question centrale concerne la description et le statut, controversés, des multiples productions glottales relevées dans la langue et qui semblent constituer un indice perceptuel pour l'identification de l'ascendance clanique d'un locuteur. Outre le cas du korebaju et d'une possible variation dialectale, un éclairage sur le statut des productions glottales en korebaju est aussi attendu par les auteurs de travaux actuels portant sur la description des processus de glottalisation dans les autres langues de la famille Tukano et également pour ce qui concerne l'établissement de leur ancêtre commun (Chacón, 2016 ; Stenzel, 2007). Pour ces raisons, nous avons inclus dans notre étude la première comparaison dialectale du korebaju en examinant les variétés korebaju et tama.

L'objectif de cette communication est (1) de présenter l'inventaire et la distribution des différents types de productions glottales relevées au cours des enquêtes linguistiques réalisées de 2022 à 2023 ; (2) de proposer leurs caractérisations acoustiques en tenant compte des variations intra- et inter-locuteurs ; (3) d'ouvrir la discussion avec des experts en traitement automatique de la parole en vue du développement d'un outil de segmentation et d'étiquetage automatique ou semi-automatique des corpus de terrain s'appuyant sur le signal acoustique et le signal électroglottographique (EGG).

La caractérisation des productions glottales en korebaju se heurte aussi à la difficulté de décrire, de manière générale, les productions glottales et leurs indices acoustiques, ainsi que leur statut dans les langues. Des discussions sont toujours très actuelles, tant sur le plan phonétique (Garellek, 2019 Garellek et al. 2023 ; Mittal et al., 2014 ; Esposito, 2012 ; Ladefoged & Maddieson, 1996), que phonologique (Dąbkowski, 2023 ; Kim & Pulleyblank, 2009 ; Macaulay & Salmons, 1995 ; Michaud, 2005 ; Nguyễn, 2021 ; Roberts, 2006 ; Stenzel, 2007). Ce débat s'inscrit aussi dans les propositions de critères pertinents pouvant permettre de classer correctement les différentes réalisations glottales. Récemment, Garellek et al. (2023) illustrent cette controverse en livrant les différentes transcriptions que ces sons ont eu à travers l'histoire, basées à la fois sur des aspects articulatoires et sur leur statut segmental ou suprasegmental pour proposer ensuite l'étiquetage de trois grand types de productions glottales pour les voyelles : « […] *(i) 'creaky/laryngealized' vowels were creaky throughout the vowel; (ii) 'rearticulated/ glottalized' vowels were creakiest in the middle of the vowel; and (iii) 'checked' vowels were creakiest towards the end of the vowel.* » (Garellek et al., 2023: 313). L'absence de consensus à propos des productions glottales concerne également la délimitation des frontières de ces segments en fonction de leurs caractéristiques acoustiques et de leurs variations (p. ex. Esposito & Khan, 2020 ; Keating et al., 2010).

Enfin, des études ont porté sur la détection et l'analyse automatiques des glottalisations à partir du signal EGG (Nguyễn, 2021 pour un dialecte de la langue Muong) sans pour autant parvenir à un résultat satisfaisant en raison de la grande variété de structures acoustiques de ces productions, parfois au sein du même segment, tandis que d'autres se sont concentrées sur la détection des cycles glottiques dans le signal de parole afin de catégoriser automatiquement le type de phonation (pour une revue, Daoudi, 2021: 92-138). Cependant la performance des outils reste bien en deçà du niveau de l'expertise humaine.

2 Corpus

Douze locuteurs de chaque variété (pour un total de 24 locuteurs) ont été enrôlés, avec leur consentement, dans cette étude : 6 hommes (H) et 6 femmes (F) de deux générations distinctes, 18-31 ans (groupe G1) et 42-70 ans (groupe G2). Une liste de 132 mots insérés chacun dans une phrase porteuse a été collectée auprès de chaque participant avec une synchronisation des signaux de la parole, de l'EGG, du flux d'air oral et du flux d'air nasal, en utilisant l'électroglottographe D800 de Laryngograph® avec une fréquence d'échantillonnage du signal de 24 kHz. Ces données ont été recueillies dans chaque communauté auprès des locuteurs de chaque variété (korebaju et tama) et à certaines heures de la journée, pour minimiser au mieux, dans nos enregistrements, l'environnement sonore de la forêt amazonienne.

La segmentation et l'étiquetage du corpus ont été réalisés manuellement pour l'ensemble des locuteurs participants. Toutes les productions glottales relevées [ʔ̣, V̰, Vˀ, ˀV, Vˀ, ˀV, ʔ] ont été labelisées à partir de l'écoute et de l'expertise visuelle du spectrogramme et du signal de parole, et en s'appuyant sur des descriptions articulatoires (Esling et al., 2019) et

acoustiques (Garellek, 2019 ; Gordon & Ladefoged, 2001; Ladefoged & Maddieson, 1996) de la littérature.

3 Panorama des glottalisations en korebaju

Le korebaju présente plusieurs types de segments glottiques allant d'une approximante glottale craquée voisée [ʔ̰] (Ladefoged & Maddieson, 1996), à des voyelles craquées (laryngalisées) [V̰], des voyelles glottalisées (réarticulées) [Vˀʔ̰], [Vˀˀ] (selon la typologie de Garellek et al. 2023 citée plus haut Section 1), et des segments correspondant à une occlusion glottale [ʔ]. Ces productions présentent toutefois des variations intra- et interlocuteur dans deux types de séquences (racine ou affixe ou racine+affixe) : (1) dans les séquences (C)VʔV, la glottalisation peut affecter les deux voyelles adjacentes telles qu'elles sont réalisées glottalisées ([Vˀʔ̰] ou [Vˀˀ] pour la première et [ʔ̰V] ou [ˀˀV] pour la seconde), et le segment intervocalique peut être produit avec différents types phonatoires allant, sur un continuum de la taille de la constriction glottique (Garellek, 2023, p. 308), de l'approximante glottale craquée voisée (+ ouverte) jusqu'à l'occlusive glottale (+ fermée) ; (2) dans les séquences (C)V̰CV ou (C)Vˀʔ̰CV ou (C)VˀˀCV (sans changement de sens), où la glottalisation est généralement une réarticulation glottale à partir de la seconde moitié de la première voyelle, mais elle peut être aussi parfois une glottalisation de toute la durée de cette voyelle (voyelle laryngalisée), et tous les timbres vocaliques sont pareillement concernés.

3.1 Approximante glottale craquée voisée [ʔ̰] et occlusion glottale [ʔ]

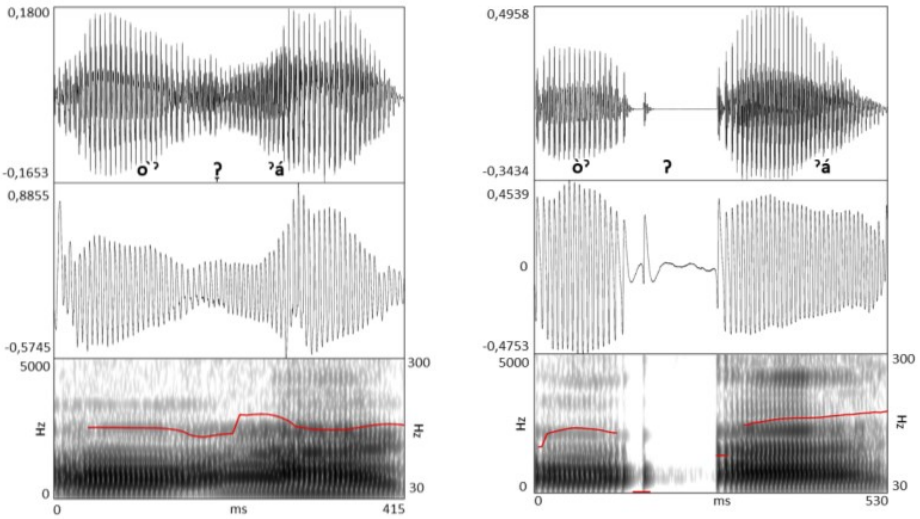


FIGURE 1 : Mot 'abeille' produit par un locuteur tama [òˀʔ̰ˀá] (à gauche) et un locuteur korebaju [òˀʔˀá] (à droite).

Dans les deux variétés korebaju, deux types de constrictions glottales sont trouvés en position intervocalique (C)V_V et leurs productions peuvent varier d'un locuteur à l'autre et parfois chez un même locuteur (FIGURE 1). Leur statut phonologique est encore en cours d'étude. Le segment intervocalique est parfois une occlusive glottale prototypique réalisée avec une fermeture complète et tenue de la glotte. Cependant, cette production semble présenter des variantes libres pouvant aller jusqu'au relâchement de la fermeture, avec oscillation des plis vocaux, et tendre vers une réalisation approximante [ʔ̞].

L'approximante glottale craquée voisée [ʔ̞], bien qu'il y ait peu d'études sur la description de ce son, est acoustiquement caractérisée par une diminution de l'énergie entre les deux voyelles, causée par une constriction partielle des plis vocaux, sans pour autant aboutir à une occlusion totale (Ladefoged & Maddieson, 1996, p. 77). Son degré de constriction est assez faible. On observe en général un abaissement de la F_0 toutefois moins important que dans les productions des voyelles craquées prototypiques (pour ces dernières au-dessous de 50 Hz chez les hommes et chez les femmes). De plus, on n'observe pas d'interruption ou d'irrégularité de la F_0 comme dans le cas des productions craquées prototypiques (laryngalisation). L'analyse en cours du signal EGG (quotient de fermeture) apportera peut-être des données physiologiques qui permettront de caractériser plus finement ce type de production.

3.2 Voyelles craquées [V̰] et voyelles réarticulées [Vʔ], [Vˀ]

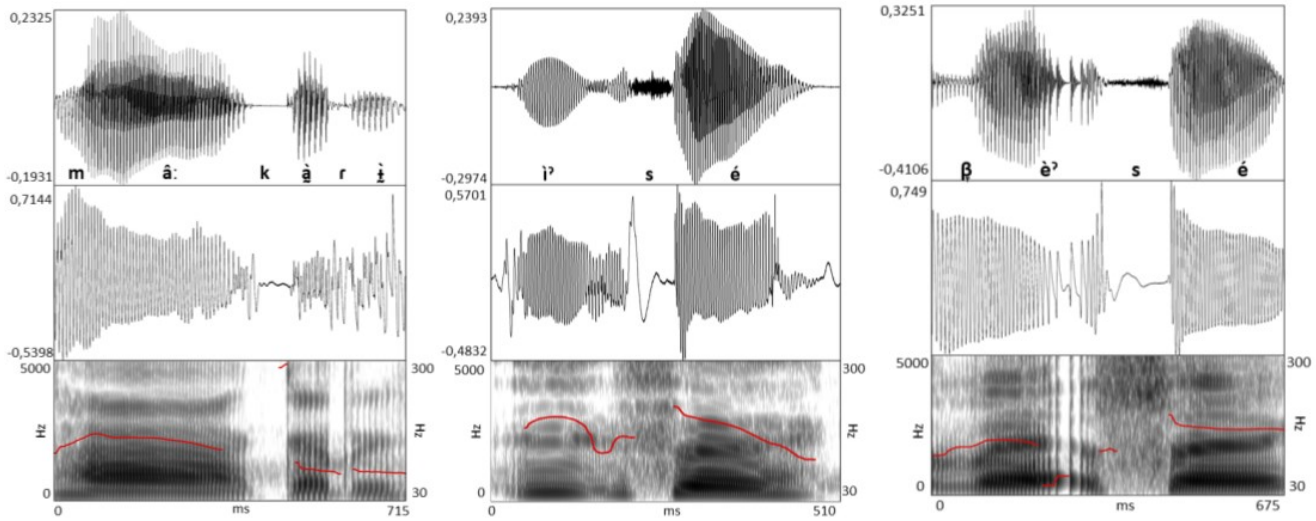


FIGURE 1 : Mot 'étroit' [mâːka̰ɾɨ̰] produit par un homme tama (à gauche), et mots 'ainsi'[ìʔsé] (au milieu) et 'sortez' [β̤èʔsé] (à droite) produits par un homme korebaju.

Les voyelles craquées prototypiques sont caractérisées par une vibration lente et irrégulière des plis vocaux, une irrégularité de la F_0 et un resserrement de la glotte, qui se traduit par

un petit pic d'ouverture, une phase de fermeture longue et un flux d'air glottique faible (Keating et al., 2015, p.1). Dans les deux variétés, les réalisations vocaliques [Vˀ], [Vˀ], [V̰] sont des allophones de /Vˀ/ (c.f. Section 1) pour lesquelles on observe une variation de craquement mais pas de timbre. Les différentes productions craquées de / Vˀ/ sont trouvées indifféremment dans des racines ou affixes ou constructions racine+affixe de type (C)VCV quel que soit le contexte consonantique (cf. FIGURE 2).

4 Conclusion

Décrire de manière fiable la phonologie du korebaju et retracer son ascendance parmi les autres langues tukano nécessite de pouvoir inventorier, caractériser et classer les productions glottales, et cela passe par la compréhension des conditions de leur réalisation. Les études acoustiques et électroglottographiques ont certainement un rôle à jouer en livrant des paramètres ou des jeux de paramètres déterminants pour pouvoir établir une typologie des productions non modales comme le montrent les travaux récents de Garellek et al. (2023) et Keating et al. (2023). La forte variabilité des types de constrictions glottiques, comme nous l'avons rencontrée en korebaju, nécessite de collecter et traiter un grand nombre de données multilocuteur avant de pouvoir poser les bases d'une description des unités sonores de la langue. Pour l'heure, la segmentation et l'étiquetage des données ne peut être réalisée de manière fiable que manuellement avec une expertise phonétique humaine. Le développement d'outils de traitement automatique des productions non modales fournirait une grande aide pour la description et la typologie des différents modes phonatoires.

Remerciements

Cette recherche bénéficie du soutien financier de l'École Doctorale 50 LLSH – UGA (mobilité sortante), de *Endangered Language Documentation Program* (ELDP Small Grant Project SG0703) et de *International Phonetic Association* (lPA). Les auteures expriment aussi leur vive reconnaissance aux membres des communautés Korebaju et Tama pour leur accueil, leur collaboration et leur participation à cette recherche.

Références

CHACÓN T. (2016). The Reconstruction of Laryngealization in Proto-Tukanoan. In H. AVELINO, M. COLER, & L. WETZELS, Éds., *The Phonetics and Phonology of Laryngeal Features in Native American Languages,* p. 258-283. Leyde : BRILL. https://doi.org/10.1163/9789004303218

COOK, D., & CRISWELL, L. (1993). *El idioma koreguaje (Tucano Occidental)*. Bogotá : Asociación Instituto Lingüístico de Verano. https://www.sil.org/resources/archives/18776

DĄBKOWSKI, M. (2023). Two grammars of A'ingae glottalization: A case for cophonologies by phase. *Natural Language & Linguistic Theory*, 1-55. https://doi.org/10.1007/s11049-023-09574-5

DAOUDI, K. (2021). *Novel paradigms in the processing of speech and its disorders*. Computer Science [cs]. Université de Bordeaux, 2021. ⟨tel-03884101⟩
ESLING, J. H., MOISIK, S. R., BENNER, A., & CREVIER-BUCHMAN, L. (2019). *Voice Quality: The Laryngeal Articulator Model*. Cambridge University Press. DOI : 10.1017/9781108696555
ESPOSITO, C. M. (2012). An acoustic and electroglottographic study of White Hmong tone and phonation. *Journal of Phonetics*, *40*(3), 466-476. https://doi.org/10.1016/j.wocn.2012.02.007
GARELLEK, M. (2019). The phonetics of voice 1. In W. F. Katz & P. F. Assmann, Éds.), *The Routledge Handbook of Phonetics*, 1re éd., p. 75-106. London : Routledge. DOI : 10.4324/9780429056253-5
GARELLEK, M., CHAI, Y., HUANG, Y., & VAN DOREN, M. (2023). Voicing of glottal consonants and non-modal vowels. *Journal of the International Phonetic Association*, *53*(2), 305-332. https://doi.org/10.1017/S0025100321000116
GORDON, M., & LADEFOGED, P. (2001). Phonation types: a cross-linguistic overview. *Journal of Phonetics*, 29, 383-406.
KEATING, P., GARELLEK, M., & KREIMAN, J. (2015). Acoustic properties of different kinds of creaky voice. *In The Scottish Consortium for ICPhS 2015, Éds., Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: the University of Glasgow. ISBN 978-0-85261-941-4. Paper number 0821.1-9 retrieved from http://www.internationalphoneticassociation.org/icphs/proceedings/ICPhS2015/Papers/ICPHS0821.pdf
KEATING, P., KUANG, J., GARELLEK, M., ESPOSITO, C. (2023). A cross-language acoustic space for vocalic phonation distinctions. *Language*, *99*, 351-389. https://doi.org/10.1353/lan.2023.a900607
KIM, E.-S., & PULLEYBLANK, D. (2009). Glottalization and Lenition in Nuu-chah-nulth. *Linguistic Inquiry*, *40*(4), 567-617. https://doi.org/10.1162/ling.2009.40.4.567
LADEFOGED, P., & MADDIESON, I. (1996). *The Sounds of the World's Languages*. Oxford : Blackwell.
MACAULAY, M., & SALMONS, J. C. (1995). The Phonology of Glottalization in Mixtec. *International Journal of American Linguistics*, *61*(1), 38-61. https://doi.org/10.1086/466244
MICHAUD, A. (2005). Final Consonants and Glottalization: New Perspectives from Hanoi Vietnamese. *Phonetica*, *61*(2-3), 119-146. https://doi.org/10.1159/000082560
MITTAL, V. K., Yegnanarayana, B., & Bhaskararao, P. (2014). Study of the effects of vocal tract constriction on glottal vibration. *The Journal of the Acoustical Society of America*, *136*(4), 1932-1941. https://doi.org/10.1121/1.4894789
MOSELEY, C. Éds,. (2010). *Atlas of the world's languages in danger,* (3e éd.). UNESCO.
NGUYỄN, M.-C. (2021). *Glottalization, tonal contrasts and intonation: An experimental study of the Kim Thuong dialect of Muong*. Doctoral dissertation. Université Sorbonne Nouvelle.
ROBERT, J. (2006). As old becomes new: glottalization in Vermont. *American Speech*, *81*(3), 227-249. https://doi.org/10.1215/00031283-2006-016
STENZEL, K. (2007). Glottalization and Other Suprasegmental Features in Wanano. *International Journal of American Linguistics*, *73*(3), 331-366. https://doi.org/10.1086/521730
VEGA RODRIGUEZ, J., VALLÉE, N., CHACÓN, T., SAVARIAUX, C., & GERBER, S. (2023). *An Intra- and Inter-Dialectal Study of Korebaju Vowels*. In *2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages* (SIGUL 2023), p. 18-20, August 2023, Dublin, Ireland. https://sigul-2023.ilc.cnr.it/wp-content/uploads/2023/08/17_Paper.pdf
VEGA RODRIGUEZ, J. (2019). The Vowel System of Korebaju. In *Proceedings of Interspeech 2019*, p. 3975-3979, doi: 10.21437/Interspeech.2019-3210

Actes des journées LIFT 2023
Nancy, France, 20-21 novembre 2023

Transfert *zero-shot* pour l'étiquetage morphosyntaxique : analyse de l'impact de la transformation des données à étiqueter pour les dialectes alsaciens

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

RÉSUMÉ

Nous présentons et évaluons une méthode de transformation des données pour améliorer les performances du transfert *zero-shot* pour l'étiquetage morphosyntaxique des dialectes alsaciens. Le corpus à annoter est transformé à l'aide de trois procédures simples, reposant notamment sur des lexiques bilingues alsacien-allemand. Les résultats obtenus avec des modèles entraînés pour l'étiquetage morphosyntaxique en diverses langues (de Vries *et al.*, 2022), plus ou moins proches de l'alsacien, montrent des gains importants sans ré-entraînement.

ABSTRACT

Zero-shot Transfer for POS Tagging : an Analysis of the Impact of Target Data Transformations for the Alsatian Dialects

We present and evaluate a data transformation method to improve the performance of *zero-shot* transfer for POS tagging of Alsatian dialects. The corpus to be annotated is transformed using three simple procedures, based in particular on bilingual Alsatian-German lexicons. The results obtained with models trained for POS tagging in various languages (de Vries *et al.*, 2022), more or less similar to Alsatian, show substantial gains without retraining.

MOTS-CLÉS : dialectes alsaciens ; transfert *zero-shot* ; étiquetage morphosyntaxique.

KEYWORDS: Alsatian dialects ; zero-shot transfer ; POS tagging.

1 Introduction

Les grands modèles de langue multilingues se prêtent aux approches par transfert ne nécessitant pas de ressources annotées pour la langue cible (*zero-shot*). Ces approches sont supposées être particulièrement utiles pour les langues disposant de peu de ressources ; nous nous intéresserons ici plus particulièrement à l'étiquetage morphosyntaxique pour les dialectes alsaciens.

Les travaux de de Vries *et al.* (2022) ont montré que différents éléments sont à prendre

en compte pour l'étiquetage morphosyntaxique par transfert inter-langues et notamment la présence de textes de la langue cible dans le corpus utilisé pour le pré-entraînement du modèle de langue. Or, les modèles multilingues les plus utilisés que sont mBERT (Devlin *et al.*, 2019; Devlin, 2019) et XLM-R (Conneau *et al.*, 2020) ont été pré-entraînés pour 104 et 100 langues, respectivement. Ainsi, la grande majorité des langues sont absentes des données utilisées pour produire ces modèles. Ceci peut expliquer en partie pourquoi la performance mesurée par de Vries *et al.* (2022) pour 65 langues sources et 105 langues cibles se limite à une exactitude moyenne de 57,4%, ce qui est faible pour une tâche réputée "simple" comme l'étiquetage morphosyntaxique.

Afin de contrebalancer l'absence de pré-entraînement pour la langue cible, plusieurs approches ont été proposées, de manière à tirer parti des ressemblances entre langues. Elles consistent à transformer des données d'une langue connue du modèle pour les rapprocher au mieux de la langue cible ou, au contraire, transformer les données de la langue cible pour les approcher d'une langue du modèle. Dans cet article, nous évaluons ce dernier type d'approche, en utilisant notamment des lexiques bilingues alsacien-allemand.

2 Transformation des données pour les méthodes par transfert

En transfert inter-langues sans ressources, la transformation des données peut se faire à différents niveaux : corpus de pré-entraînement du modèle de langue, données d'affinage pour la tâche cible, données cibles.

Ainsi, Hana *et al.* (2011) décrivent une méthode d'étiquetage morphosyntaxique pour le vieux tchèque dont la stratégie consiste à transformer un corpus de tchèque moderne pour qu'il ressemble au vieux tchèque et, inversement, à transformer un corpus de vieux tchèque pour qu'il ressemble au tchèque moderne. Ces transformations font notamment appel à des règles de modification phonémique et graphémique. Bernhard & Ligozat (2013) reprennent certaines de ces idées pour l'étiquetage morphosyntaxique de l'alsacien, en remplaçant les règles de changement par un lexique bilingue allemand-alsacien limité aux mots grammaticaux, qui sont les plus fréquents. Ce lexique est utilisé pour transposer les textes cibles en alsacien vers l'allemand, avant des les étiqueter avec des outils entraînés pour l'allemand. Cette méthode simple augmente les performances de l'étiquetage.

Wang *et al.* (2022) proposent également l'utilisation de lexiques bilingues pour synthétiser des données, en justifiant cette approche par la plus grande disponibilité de lexiques bilingues par rapport aux corpus monolingues de grande taille pour une large majorité des langues de la planète. Les lexiques sont utilisés pour traduire des mots d'une langue bien dotée vers la langue cible, afin de générer des données permettant de poursuivre le pré-entraînement du modèle de langue multilingue ou encore des données annotées pour l'affinage pour la tâche cible. Les résultats montrent une augmentation significative de la performance pour les tâches d'étiquetage morphosyntaxique, analyse syntaxique et reconnaissance d'entités

nommées.

La méthode consistant à générer des données monolingues synthétiques à l'aide d'un lexique bilingue est également utilisée par Lothritz *et al.* (2022). Leur objectif est d'entraîner un modèle BERT pour le luxembourgeois en augmentant les données par des textes partiellement traduits de l'allemand vers le luxembourgeois. Cette traduction se limite aux mots outils. Toujours pour le luxembourgeois, Song *et al.* (2023) produisent un "pseudo" corpus parallèle luxembourgeois-anglais à partir d'un corpus allemand-anglais en utilisant lexique bilingue Le pseudo-corpus luxembourgeois-anglais est ensuite utilisé pour entraîner un modèle de traduction. dont les résultats restent toutefois largement inférieurs au modèle multilingue NLLB-large (Costa-jussà *et al.*, 2022).

D'autres approches font totalement l'impasse sur les lexiques bilingues et génèrent des données synthétiques par l'injection aléatoire de bruit dans les données disponibles pour la langue source mieux dotée. Aepli & Sennrich (2022) augmentent les données de pré-entraînement du modèle de langue en injectant du bruit au niveau des caractères (suppression, insertion, remplacement) afin de générer de la variation orthographique. Dans la mesure où les mots sont découpés en sous-mots par les tokéniseurs des modèles de langue, ces perturbations conduisent à des modifications dans la segmentation des mots. Cette méthode permet d'obtenir une augmentation de l'exactitude de 22 points de pourcentage pour l'étiquetage morpho-syntaxique des dialectes suisses allemands, par rapport à une méthode sans injection de bruit utilisant uniquement des données en allemand. Blaschke *et al.* (2023) reprennent cette méthode et en font une analyse détaillée pour l'étiquetage morphosyntaxique de 7 langues appartenant à 3 familles linguistiques, incluant l'alsacien. Leur étude montre que la différence entre les proportions de mots qui ont été découpés en sous-mots dans les données source et cible a une corrélation négative avec la performance : plus cette différence est faible, plus l'exactitude est élevée. Les expériences décrites dans notre article visent à comparer, pour l'alsacien, l'approche de Blaschke *et al.* (2023), qui suppose d'entraîner un nouveau modèle à partir des données transformées, à une approche *zero-shot* par transformation des données cibles.

3 Méthode

Nous utilisons un corpus de textes alsaciens étiqueté en parties du discours selon les catégories Universal Dependencies (De Marneffe *et al.*, 2021) et comportant 12 582 tokens de surface et 12 907 mots syntaxiques (Bernhard *et al.*, 2018, 2023). Ce corpus est transformé de manière à s'approcher de l'allemand à l'aide de trois procédures simples :

1. Accentuation (`A`) : suppression des diacritiques de voyelles spécifiques aux dialectes alsaciens et conversion vers la forme non accentuée. Seuls les umlauts ⟨ä, ö, ü⟩ sont conservés, car ils sont utilisés en allemand. Les apostrophes sont également normalisées vers la forme ⟨'⟩.

2. Classes fermées (C) : utilisation d'un lexique de conversion de l'alsacien vers l'allemand de formes appartenant aux classes fermées. Nous réutilisons directement le lexique constitué par Bernhard & Ligozat (2013), sans modification. Ce lexique contient 133 entrées et a été constitué par étude d'un petit corpus de 5 textes. Un seul de ces textes se trouve également dans le corpus annoté utilisé pour l'évaluation (soit 396 tokens)
3. Classes ouvertes (O) : utilisation d'un lexique de conversion de l'alsacien vers l'allemand de formes appartenant aux classes ouvertes. Nous réutilisons le lexique produit par Bernhard (2014, 2021). Si un mot en alsacien a plusieurs traductions possibles, seule la plus fréquente est conservée (fréquence dans le corpus `deu_news_2022_1M`[1] (Goldhahn *et al.*, 2012)). Le lexique final comporte 6 699 paires de mots alsacien-allemand.

Ces procédures peuvent également être combinées entre elles, de manière à augmenter le nombre de transformations sur le corpus d'entrée. La Table 1 récapitule le nombre et le pourcentage de mots transformés par chaque procédure tandis que la Table 2 donne des exemples. Nous indiquons également le nombre moyen de sous-mots par mot après tokénisation avec XLM-R-base (Conneau *et al.*, 2020), qui est le modèle de langue utilisé pour l'entraînement des modèles d'étiquetage morphosyntaxique de de Vries *et al.* (2022) que nous utilisons dans nos expériences. Ces modèles, qui ont été entraînés pour 65 langues sources, sont ensuite appliqués pour étiqueter les différentes versions du corpus alsacien.

Toutes les expériences et analyses sont réalisées à l'aide des principaux outils et bibliothèques Python suivants : *Hugging Face*[2] pour les modèles[3], les bibliothèques *Transformers v. 4.30.2* et *Datasets v. 2.13.0* (Tunstall *et al.*, 2022), *PyTorch v. 2.0.1*[4], *pandas v. 2.0.3* (pandas development team, 2023), *scikit-learn v. 1.3.0* (Pedregosa *et al.*, 2011), *matplotlib v. 3.7.2* (Hunter, 2007) et *seaborn v. 0.12.2* (Waskom, 2021).

4 Résultats

La Table 3 détaille les résultats, en terme d'exactitude, pour les 10 langues sources qui obtiennent les meilleurs résultats en moyenne pour l'alsacien et pour les différentes transformations. A titre de comparaison, nous faisons également figurer les résultats obtenus pour les dialectes suisses allemands (Aepli & Clematide, 2018) avec les mêmes langues sources (sans transformation des données) : ces dialectes sont très proches des dialectes alsaciens, en particulier l'aire haut alémanique au sud de l'Alsace, en zone limitrophe de la Suisse.

Ces résultats montrent un impact important des transformations simples sur les résultats

1. https://wortschatz.uni-leipzig.de/en/download/German#deu_news_2022
2. https://huggingface.co
3. https://huggingface.co/wietsedv/xlm-roberta-base-ft-udpos28-[codedelangu
4. https://pytorch.org/

Traitement	# transformations	% mots	# sous-mots / mot
Aucun	0	0%	1,92
A	2 586	20%	1,70
C	2 475	19%	1,82
O	749	6%	1,88
AC	4 344	34%	1,66
AO	3 108	24%	1,68
CO	3 127	24%	1,78
ACO	4 804	37%	1,64

TABLE 1 – Nombre et pourcentage de mots syntaxiques transformés selon chaque procédure. La dernière colonne indique le nombre moyen de sous-mots par mot après tokénisation avec XLM-R-base.

Texte original	Mìt	dr	Jugend	ìsch	nit	loos	!
sous-mots	M␣ì␣t	dr	Jugend	ì␣sch	nit	loo␣s	!
A	Mit	dr	Jugend	isch	nit	loo␣s	!
C	Mit	der	Jugend	ist	nicht	loo␣s	!
O	M␣ì␣t	dr	Jugend	ì␣sch	nada	loo␣s	!
AC	Mit	der	Jugend	ist	nicht	loo␣s	!
AO	Mit	dr	Jugend	isch	nada	loo␣s	!
CO	Mit	der	Jugend	ist	nada	loo␣s	!
ACO	Mit	der	Jugend	ist	nada	loo␣s	!

TABLE 2 – Exemples de sous-mots en fonction des pré-traitements.

obtenus. La suppression des accents à elle seule permet de gagner 7,2 points d'exactitude en moyenne par rapport aux données brutes pour l'ensemble des 65 langues sources, soit plus que l'utilisation d'un lexique de mots de classes ouvertes (gain de 2,1 points en moyenne). C'est d'ailleurs cette dernière ressource qui a l'impact le plus faible sur les résultats. Le lexique de mots grammaticaux permet d'augmenter le score d'exactitude de 14,1 points en moyenne sur l'ensemble des langues sources, confirmant ainsi les observations de Bernhard & Ligozat (2013). Enfin, les meilleurs résultats sont obtenus par combinaison de l'ensemble des ressources (ACO) : exactitude moyenne de 61,0 (+18,7) pour les 65 langues sources, et de 72,1 (+21,7) pour les 10 langues présentées dans la Table 3.

Si l'on met en regard les données des Tables 1 et 3, on constate que, globalement, plus le pourcentage de mots transformés augmente, plus le nombre moyen de sous-mots par mot diminue et plus le score d'exactitude augmente également. La diminution du nombre moyen de sous-mots par mot indique, indirectement, que les données se rapprochent davantage de celles utilisées pour pré-entraîner le tokéniseur du modèle de langue.

Ces résultats sont inférieurs mais très proches du score de 78 % d'exactitude obtenu par

Langue source	suisse	alsacien	A	C	O	AC	AO	CO	ACO
afrikaans	55,2	53,0	61,0	68,6	55,1	71,6	62,2	69,3	72,0
allemand	50,2	49,9	58,7	69,5	52,5	73,0	60,4	70,6	73,6
arménien	46,2	47,6	58,9	65,8	51,2	71,1	61,2	67,8	72,2
arménien occidental	58,2	55,6	65,6	71,0	59,0	74,8	67,7	72,4	75,5
bulgare	50,3	48,5	57,8	66,0	51,5	69,8	59,9	67,4	70,8
féroïen	54,1	50,5	59,5	67,5	52,9	71,3	61,2	68,6	72,0
gallois	49,9	50,4	57,6	66,9	52,4	69,5	58,8	67,8	70,0
lituanien	49,7	48,8	57,9	66,4	51,4	69,5	59,4	67,8	70,5
roumain	53,0	51,8	60,6	69,2	54,9	73,2	62,8	70,4	74,0
tchèque	50,8	49,6	58,1	67,9	52,1	71,3	59,4	69,4	72,2

TABLE 3 – Exactitude (en %) pour différents prétraitements. Les 10 langues sources représentées sont celles qui obtiennent les meilleurs résultats en moyenne pour l'alsacien. Les langues présentes dans les données de pré-entraînement pour XLM-R sont soulignées.

Blaschke *et al.* (2023) pour le même corpus alsacien, en manipulant le corpus allemand avant l'entraînement du modèle d'étiquetage morphosyntaxique. Dans notre cas, nous n'avons pas ré-entraîné les modèles et les avons utilisés tels quels.

5 Discussion

Conclusion Nous avons analysé une méthode peu coûteuse, simple à mettre en œuvre et ne nécessitant pas de ré-entraînement de modèles pour le transfert inter-langues de modèles d'étiquetage morphosyntaxique. Les ressources requises sont limitées et peuvent être facilement constituées par une étude de corpus ou à partir d'un lexique bilingue, même de taille limitée. Les résultats obtenus pour l'étiquetage de l'alsacien montrent que tous les modèles bénéficient des transformations, pas uniquement le modèle affiné pour l'allemand.

Perspectives Le lexique des mots de classes fermées gagnerait à être étendu, ce qui pourrait contribuer à augmenter encore l'exactitude. Par ailleurs, il serait utile de comprendre et expliquer pourquoi de si bons résultats sont obtenus avec l'arménien occidental et le roumain. Les bonnes performances globales du roumain, pour un large ensemble de langues cibles, avait déjà été remarqué par de Vries *et al.* (2022).

Limites Les travaux présentés dans cet article ont été réalisés pour une seule langue cible. Par ailleurs, le corpus alsacien utilisé ne représente qu'une partie de la variation observée dans l'espace dialectal germanique en Alsace et ne saurait être représentatif de l'ensemble des locutrices et locuteurs. Il faudrait donc étendre les expériences à d'autres langues pour vérifier si les conclusions restent valides. Enfin, une seule tâche a été évaluée (classification de tokens, et, plus particulièrement, étiquetage morphosyntaxique) et il faudrait donc vérifier si les transformations proposées ont un impact similaire pour d'autres types de tâches.

Remerciements

Nous remercions le Centre de Calcul Haute Performance de l'Université de Strasbourg pour avoir soutenu ce travail en fournissant un support scientifique et l'accès aux ressources informatiques. Une partie des ressources informatiques a été financée par le projet Equipex Equip@Meso (Programme Investissements d'Avenir) et le CPER Alsacalcul/Big Data.

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

Références

AEPLI N. & CLEMATIDE S. (2018). Parsing Approaches for Swiss German. In *Proceedings of SwissText 2018*.

AEPLI N. & SENNRICH R. (2022). Improving Zero-Shot Cross-lingual Transfer Between Closely Related Languages by Injecting Character-Level Noise. In *Findings of the Association for Computational Linguistics : ACL 2022*, p. 4074–4083, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-acl.321.

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

BERNHARD D. (2021). Lexique multilingue alsacien – français – allemand relié aux synsets de BabelNet. DOI : 10.34847/nkl.3f9b2i11.

BERNHARD D., ERHART P., HUCK D. & STEIBLÉ L. (2023). Annotated Corpus for the Alsatian Dialects. version 3.0, DOI : 10.5281/zenodo.1170128.

BERNHARD D. & LIGOZAT A.-L. (2013). Es esch fàscht wie Ditsch, oder net ? Étiquetage morphosyntaxique de l'alsacien en passant par l'allemand. In *Actes de TALARE 2013 : Traitement Automatique des Langues Régionales de France et d'Europe*, p. 209–220, Les Sables d'Olonne, France.

BERNHARD D., LIGOZAT A.-L., MARTIN F., BRAS M., MAGISTRY P., VERGEZ-COURET M., STEIBLE L., ERHART P., HATHOUT N., HUCK D., REY C., REYNÉS P., ROSSET S., SIBILLE J. & LAVERGNE T. (2018). Corpora with Part-of-Speech Annotations for Three Regional Languages of France : Alsatian, Occitan and Picard. In N. CALZOLARI, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, K. HASIDA, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK, S. PIPERIDIS & T. TOKUNAGA, Éds., *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference*, p. 3917–3924, Miyazaki, Japan.

BLASCHKE V., SCHÜTZE H. & PLANK B. (2023). Does Manipulating Tokenization Aid Cross-Lingual Transfer ? A Study on POS Tagging for Non-Standardized Languages. In

Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), p. 40–54, Dubrovnik, Croatia.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 8440–8451, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.747.

COSTA-JUSSÀ M. R., CROSS J., ÇELEBI O., ELBAYAD M., HEAFIELD K., HEFFERNAN K., KALBASSI E., LAM J., LICHT D. & MAILLARD J. (2022). No language left behind : Scaling human-centered machine translation. arXiv : 2207.04672.

DE MARNEFFE M.-C., MANNING C. D., NIVRE J. & ZEMAN D. (2021). Universal dependencies. *Computational linguistics*, **47**(2), 255–308.

DE VRIES W., WIELING M. & NISSIM M. (2022). Make the Best of Cross-lingual Transfer : Evidence from POS Tagging with over 100 Languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 7676–7685, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.529.

DEVLIN J. (2019). Multilingual bert readme document. https://github.com/google-research/bert/blob/master/multilingual.md.

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

GOLDHAHN D., ECKART T. & QUASTHOFF U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection : From 100 to 200 Languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, p. 759–765, Istanbul, Turkey : European Language Resources Association (ELRA).

HANA J., FELDMAN A. & AHARODNIK K. (2011). A low-budget tagger for Old Czech. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, p. 10–18.

HUNTER J. D. (2007). Matplotlib : A 2d graphics environment. *Computing in Science & Engineering*, **9**(3), 90–95. DOI : 10.1109/MCSE.2007.55.

LOTHRITZ C., LEBICHOT B., ALLIX K., VEIBER L., BISSYANDE TEGAWENDE., KLEIN J., BOYTSOV A., LEFEBVRE C. & GOUJON A. (2022). LuxemBERT : Simple and practical data augmentation in language model pre-training for luxembourgish. In *Proceedings of the Language Resources and Evaluation Conference*, p. 5080–5089, Marseille, France : European Language Resources Association.

PANDAS DEVELOPMENT TEAM (2023). pandas-dev/pandas : Pandas v2.0.3. 10.5281/zenodo.8092754, DOI : 10.5281/zenodo.8092754.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

SONG Y., EZZINI S., KLEIN J., BISSYANDE T., LEFEBVRE C. & GOUJON A. (2023). Letz Translate : Low-Resource Machine Translation for Luxembourgish. In *5th International Conference on Natural Language Processing*, Guangzhou, China.

TUNSTALL L., VON WERRA L. & WOLF T. (2022). *Natural Language Processing with Transformers*. O'Reilly Media, Inc.

WANG X., RUDER S. & NEUBIG G. (2022). Expanding Pretrained Models to Thousands More Languages via Lexicon-based Adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 863–877, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.61.

WASKOM M. L. (2021). seaborn : statistical data visualization. *Journal of Open Source Software*, **6**(60), 3021. DOI : 10.21105/joss.03021.

Deuxième partie

Session dédiée aux approches formelles

Actes des journées LIFT 2023
Nancy, France, 20-21 novembre 2023

A Layered Approach to Semantic Representation

Siyana Pavlova Maxime Amblard Bruno Guillaume
Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
{firstname.lastname}@loria.fr

RÉSUMÉ

Une approche stratifiée pour la représentation sémantique

Nous présentons une première version de notre approche de la représentation sémantique. Bien qu'il existe de nombreux formalismes pour la sémantique, parvenir à rassembler plusieurs informations sémantiques tout en restant lisible reste une tâche difficile. Nous proposons de nous appuyer sur la structure prédicat-argument (PA) utilisée dans les AMR et d'étendre la représentation avec de nouveaux type de sommets et d'arêtes. Nous introduisons la notion de "caractéristiques" qui représente des phénomènes sémantiques nouvellement encodés sous forme de couches. Des arêtes font le lien avec les caractéristiques sémantiques et la structure PA, ou dans le cas de phénomènes sémantiques en interaction avec d'autres caractéristiques. Notre approche permet de conserver la traditionnelle structure PA et de se focaliser sur des phénomènes particuliers et leurs interactions. Un avantage explicite est aussi de permettre de rendre compte de phénomènes complexes qui utilise la portée (négation, quantification, etc.).

ABSTRACT

In this article, we present a first version of a layered approach we take to semantic representation. Our approach is motivated by the need for a semantic representation formalism which can encode a rich variety of semantic phenomena, while remaining simple to annotate and easy to read. Our representation derives its core predicate-argument structure from Abstract Meaning Representation (AMR), but is then extended with a number of types of nodes and edges. We introduce "features" - nodes for each semantic phenomenon we wish to encode. Features represent various layers of our representation. Edges are then attached to each feature node, that can link to nodes from the predicate-argument structure of the representation or, in the case of interacting semantic phenomena, to edges from other layers. An advantage of our approach is the possibility to exclude layers from the representation easily, while still being able to represent phenomena such as scope that are difficult for most graph-oriented formalisms.

MOTS-CLÉS : formalismes de représentation sémantique.

KEYWORDS: semantic representation formalisms, layered semantic representation.

1 Introduction

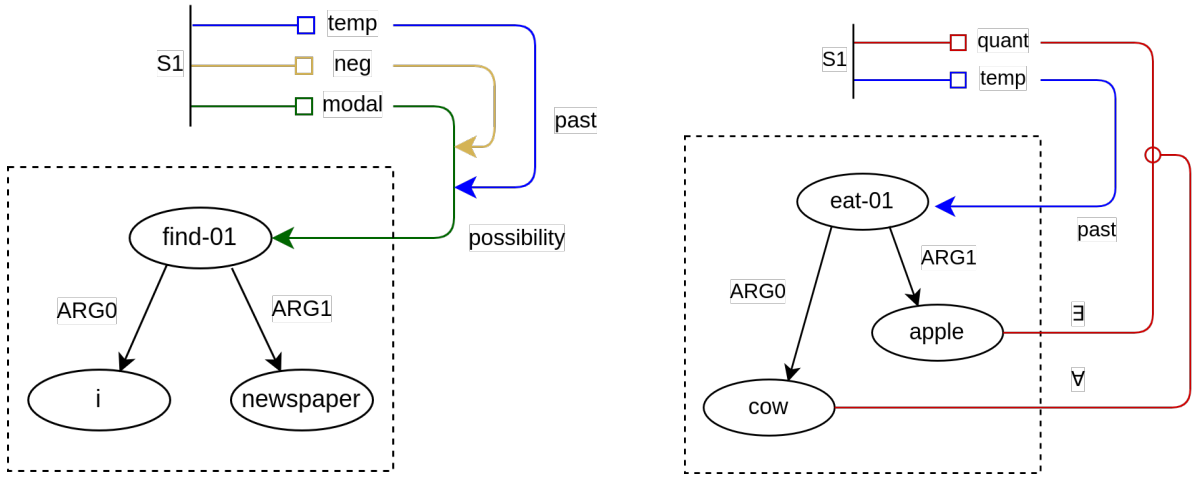
Current semantic representation formalisms can be compared across multiple aspects, as shown in the literature (Abend & Rappoport, 2017; Žabokrtský *et al.*, 2020; Pavlova *et al.*, 2023b). In this work, we want to focus on the distinction in their ability to encode various semantic phenomena, and their readability. Thus, we split formalisms into two broad categories - logic-based (Kamp & Reyle, 1993; Montague, 1970), and graph-based (Banarescu *et al.*, 2013; Abend & Rappoport, 2013; White *et al.*, 2016). Logic-based formalisms tend to be powerful in terms of encoding, being able to express phenomena such as quantifier scope, but are not easy to read and interpret without prior training in Logic. Graph-based ones, on the other hand, are easier to read and annotate, but often lack when it comes to expressing scope or being compositional. Admittedly, some graph-based formalisms, such Minimal Recursion Semantics (MRS) (Copestake *et al.*, 2005) are able to encode a large set of phenomena. However, the way this is realised still makes the formalism difficult to read. Similarly, logic formulas can be represented as graphs. However, interpreting those still requires an understanding of Logic.

Many of the more recent graph-based formalisms have been proposed with large-scale annotation in mind, Abstract Meaning Representation (AMR) (Banarescu *et al.*, 2013) being the most popular one. However, to achieve this, they do not annotate various phenomena, such as temporal information, definiteness, plurality, etc. Uniform Meaning Representation (Van Gysel *et al.*, 2021), which can be seen as a standalone formalism, puts together a number of AMR extensions, achieving a rather more powerful representation than AMR alone. However, as can be seen in Figure 1 of (Zhao *et al.*, 2021), since all semantic phenomena are added as nodes and edges of the same status as those in the original AMR structure, the representation becomes rather messy for longer texts.

Our goal with this work is to create a formalism which is a common ground between logic-based and graph-based ones - easy to read and annotate, but also able to encode multiple semantic phenomena. Furthermore, we want to keep each phenomenon separate from the central predicate-argument structure in order to keep the overall structure more organised, and to allow for the independent exploration of different phenomena. Thus, we propose a way to modify and extend an AMR-inspired graph representation by adding explicit layers to the structure. Each layer is designed to encode a different semantic phenomenon, allowing a straightforward way to "switch off" the layer. This avoids cluttering of the representation when certain phenomena are not the focus of a given task.

2 Proposal

We present here our proposal to extend graph structures in order to encode various semantic phenomena in specifically dedicated layers. For this purpose, we add a number of node types



(a) "I could not find the newspaper." (b) "Every cow ate an apple."

Figure 1: Our representation for two sentences

and relation types. We will explore these with the help of the two examples in English shown in Figure 1: for the sentence "*I could not find the newspaper*" (Figure 1a) with negation and modality, and "*Every cow ate an apple*" (Figure 1b), with the usual scope ambiguity.

The predicate-argument structure of each sentence is represented as a standard graph (the part inside the dashed rectangle), with nodes (from a set of nodes $\boldsymbol{V}$) and edges (from a set of edges $\boldsymbol{E}$). The nodes are predicates and concepts. The edges are argument roles, and are labeled. The predicates and the argument roles have been taken from PropBank (Palmer *et al.*, 2005). This is the standard structure we see in AMR and AMR-derived formalisms.

However, we restrict the usual AMR notation to the predicate-argument structure of the main predicate of a phrase. By doing this we assume to follow a neo-davidsonian representation of semantics (Davidson, 1967; Parsons, 1990). In this line, we extend this by first adding a node $\boldsymbol{S}$ to represent the event itself (`S1` in the two figures). For multiple events in the same sentence, we introduce a node of the same type for each event, as we will see in section 3.

Then, a set of features $\boldsymbol{F}$ is added, each representing a semantic phenomenon that we wish to encode in the structure. Each feature is linked to an event node. In Figure 1a, we represent temporality, modality and negation. In Figure 1b, in addition to temporality, we also encode quantification.

Next, we introduce a set of edges, $\boldsymbol{E'}$, which link a feature to a node from $\boldsymbol{V}$, as we see with `modal` and `find-01` in Figure 1a, or with `temp` and `eat-01` in Figure 1b.

Figure 1a demonstrates that there are cases where it is not enough to link features only to nodes from $\boldsymbol{V}$. In that sentence, it is not the *finding* event that is expressed in the past, but the speaker's *ability* to do so. Thus, we introduce another set of edges, which link a feature to an edge from $\boldsymbol{E'}$. This is also useful for negation as we can see in the same sentence.

Finally, as demonstrated in Figure 1b, when more than one quantifier is present in a sentence,

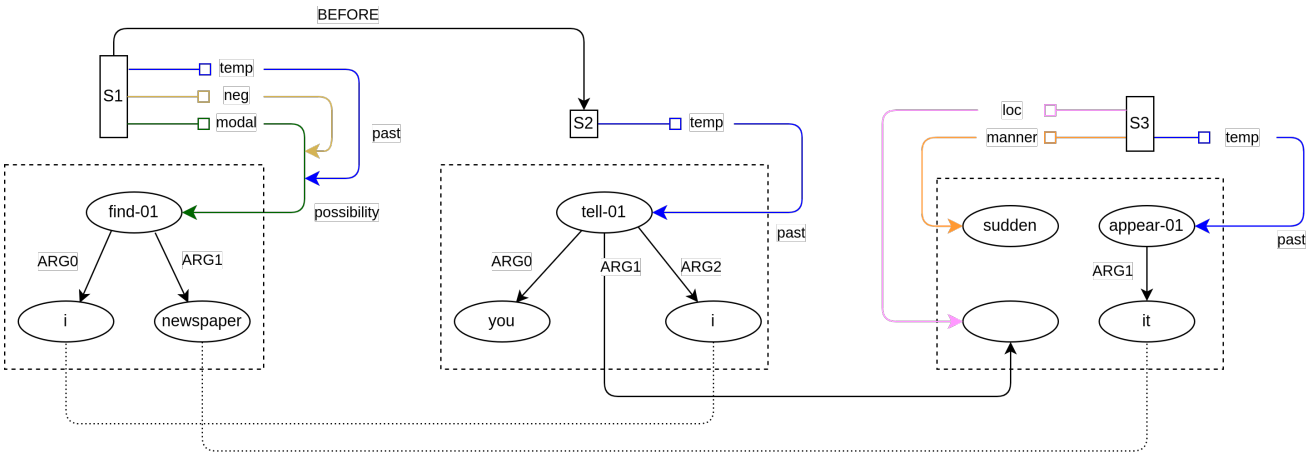


Figure 2: "I could not find the newspaper until you told me where it would suddenly appear."

we might need to specify which one takes scope over which other ones. To be able to do this, in the quantification layer we introduce another type of edges, which can link an edge from $\boldsymbol{E'}$ to a node from $\boldsymbol{V}$. Edges linking to and from features are labeled or unlabeled, depending on the feature they are linked to.

One of the advantages of the thus described structure is that while various phenomena can be encoded, it is easy to choose to not represent a layer if the phenomenon it represents is not relevant for a given task. In Figure 1a, even though it is possible, we have chosen to not represent quantification.

3 Towards more complex examples

Let us now imagine a more complex example with multiple events. In Figure 2, we have our representation for the sentence "*I could not find the newspaper until you told me where it would suddenly appear*", where we have three events: *finding*, *telling*, and *appearing*.

As pointed out earlier, when dealing with multiple events, we choose to follow a neo-davidsonian approach, treating each event separately. Thus, we introduce three event nodes, `S1` for "*find*", `S2` for "*told*" and `S3` for "*appear*". The word "*until*" expresses the fact that "*could not find*" was happening before "*told*". Thus, we employ the use of another set of edges, $\boldsymbol{E_S}$, which links event nodes. We add one such edge between `S1` and `S2` with the label `BEFORE`. Edges in $\boldsymbol{E_S}$ are labeled. They can be thought of as discourse relations, demonstrating our proposal's ability to go beyond semantics.

Using edges from $\boldsymbol{E_S}$ is one of the ways we propose for handling subordinate clauses, specifically those introduced by subordinating conjunctions such as "because", "although", "until". However, some subordinating conjunctions can be directly linked as arguments of the predicate in the main clause, as we see with the example of "*where it would suddenly appear*". Here, the location represented by "*where*" is the `ARG1` of the predicate `tell-01`

as per PropBank. Thus, while the *appearing* event is represented in its own box, one of its arguments, namely, the location, is linked to `tell-01` by an edge from **E**.

As we see in `S3`'s box, we also treat AMR's non-core arguments differently. These are optional modifiers that are not predicate-dependent. As such, we link them to the event node via a new feature, rather than to the main predicate of that event. We can see that with `sudden` being linked to `S3` via the `manner` feature in Figure 2. This helps us to keep the core predicate-argument structure separate.

Finally, we have two instances of co-reference in this example, between "*I*" and "*me*" and between "*newspaper*" and "*it*". Our current approach is to link these with a pair of dashed edges, different in type from the edges in ***E***, running between each instance of the co-referent. These edges are unlabeled.

With the example in this section, we demonstrate our proposal's versatility in encoding various phenomena even for longer texts. The use of layers allows us to represent a wide range of semantic phenomena while keeping it easy to identify each of them in the overall structure. Representing each event separately allows us to maintain this aspect of the representation even for longer texts.

4 Discussion and Future Work

In this work, we do not provide all the definitions of the framework, but as part of ongoing work we have explored its ability to encode a lot of different phenomena and how they could interact in the representation, namely: temporality, aspect, modality, negation, quantifier scope, definiteness, plurality and generics. Thus, we believe our representation has the potential to encode the phenomena represented by logic-based formalisms. However, with our proposal on how to visualise it, it is also more easy to read and annotate. Finally, the clear separation between phenomena allows for a more organised view and the possibility to "switch off" certain layers when irrelevant for a specific task.

As we have seen, edges linking to and from features can be labeled or unlabeled, depending on the feature they are connected to. The focus of this work so far has been the general structure of our proposal. As such, we have used only a simple set of classes for each feature: *possibility* and *necessity* for `modal`, *past*, *present* and *future* for `temp`. In the future, we intend to explore the lattice approach proposed in (Van Gysel *et al.*, 2019) and utilised in UMR. Furthermore, as we can see with the example in Figure 2, while the three events are all in the past, it is clear to a speaker that there is an order in which they occur - the *inability to find the newspaper* happens before the *telling*, while *the appearance of the newspaper* happens after it. This information is not captured by the current version of our proposal but will be the subject of future work, where we could utilise Reichenbach's notions of *speech time*, *event time* and *reference time* (Reichenbach, 1947).

Further next steps for this project will be to give the formal definition of our proposal, expand the list of phenomena we encode, annotate a first exploration corpus, and provide annotation guidelines. This work will be followed by transformation experiments between the thus proposed formalism and other semantic representation formalisms in the line of our previous work (Pavlova *et al.*, 2023a). Furthermore, to allow for this representation to be used for semantic parsing and tasks further up the NLP pipeline, we want to develop an equivalent textual representation in the spirit of PENMAN notation (Matthiessen & Bateman, 1991). Finally, our work so far has been guided and informed by examples in English. For future versions, we plan to take languages other than English into account in order to make sure we are designing a universal framework.

Acknowledgements

We would like to thank the anonymous reviewers for their feedback and comments. Part of this work has been funded by *Agence Nationale de la Recherche* (ANR, fr: National Research Agency), grant number ANR-20-THIA-0010-01.

References

ABEND O. & RAPPOPORT A. (2013). Universal Conceptual Cognitive Annotation (UCCA). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 228–238, Sofia, Bulgaria: Association for Computational Linguistics.

ABEND O. & RAPPOPORT A. (2017). Tthe State of the Art in Semantic Representation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 77–89, Vancouver, Canada: Association for Computational Linguistics. DOI : 10.18653/v1/P17-1008.

BANARESCU L., BONIAL C., CAI S., GEORGESCU M., GRIFFITT K., HERMJAKOB U., KNIGHT K., KOEHN P., PALMER M. & SCHNEIDER N. (2013). Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, p. 178–186, Sofia, Bulgaria: Association for Computational Linguistics.

COPESTAKE A., FLICKINGER D., POLLARD C. & SAG I. A. (2005). Minimal Recursion Semantics: An Introduction. *Research on language and computation*, **3**(2), 281–332.

DAVIDSON D. (1967). The Logical Form of Action Sentences. In N. RESCHER, Éd., *The Logic of Decision and Action*, p. 81–95. University of Pittsburgh Press.

KAMP H. & REYLE U. (1993). *From Discourse to Logic Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Dordrecht. Kluwer.

MATTHIESSEN C. M. & BATEMAN J. A. (1991). Text Generation and Systemic-Functional Linguistics: Experiences from English and Japanese. *Communication in Artificial Intelligence Series*, **19**(1).

MONTAGUE R. (1970). English as a Formal Language.

PALMER M., GILDEA D. & KINGSBURY P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, **31**(1), 71–106. DOI : 10.1162/0891201053630264.

PARSONS T. (1990). *Events in the Semantics of English: A Study in Subatomic Semantics*. MIT Press.

PAVLOVA S., AMBLARD M. & GUILLAUME B. (2023a). Bridging Semantic Frameworks: mapping DRS onto AMR. In *The 15th International Conference on Computational Semantics (IWCS 2023)*, Nancy, France. HAL : hal-04129563.

PAVLOVA S., AMBLARD M. & GUILLAUME B. (2023b). Structural and Global Features for Comparing Semantic Representation Formalisms. In *The 4th International Workshop on Designing Meaning Representation*, Nancy, France. HAL : hal-04129557.

REICHENBACH H. (1947). *Elements of Symbolic Logic*. A Free Press paperback : philosophy. Macmillan Company.

VAN GYSEL J. E., VIGUS M., CHUN J., LAI K., MOELLER S., YAO J., O'GORMAN T., COWELL A., CROFT W., HUANG C.-R. *et al.* (2021). Designing a Uniform Meaning Representation for Natural Language Processing. *KI-Künstliche Intelligenz*, **35**(3-4), 343–360.

VAN GYSEL J. E. L., VIGUS M., KALM P., LEE S.-K., REGAN M. & CROFT W. (2019). Cross-Linguistic Semantic Annotation: Reconciling the Language-Specific and the Universal. In *Proceedings of the First International Workshop on Designing Meaning Representations*, p. 1–14, Florence, Italy: Association for Computational Linguistics. DOI : 10.18653/v1/W19-3301.

WHITE A. S., REISINGER D., SAKAGUCHI K., VIEIRA T., ZHANG S., RUDINGER R., RAWLINS K. & VAN DURME B. (2016). Universal Decompositional Semantics on Universal Dependencies. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 1713–1723.

ŽABOKRTSKÝ Z., ZEMAN D. & ŠEVČÍKOVÁ M. (2020). Sentence Meaning Representations Across Languages: What Can We Learn from Existing Frameworks? *Computational Linguistics*, **46**(3), 605–665. DOI : 10.1162/coli_a_00385.

ZHAO J., XUE N., VAN GYSEL J. & CHOI J. D. (2021). UMR-Writer: A Web Application for Annotating Uniform Meaning Representations. In H. ADEL & S. SHI, Éds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, p. 160–167, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-demo.19.

Actes des journées LIFT 2023
Nancy, France, 20-21 novembre 2023

Analysing topic shifts in task-oriented dialogues

Amandine Decker[1, 2] Maxime Amblard[1] Ellen Breitholtz[2]
(1) Universite de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
(2) CLASP, University Göteborg
{amandine.decker, maxime.amblard}@univ-lorraine.fr,
ellen.breitholtz@ling.gu.se

RÉSUMÉ

Analyser les changements de sujet dans les conversations ayant un but.

Le sujet d'une conversation joue un role crucial dans sa cohérence globale car ce qui est discuté à un moment donné limite les contributions possibles des participants. Lancer un nouveau sujet de conversation alors que le précédent est encore en cours de discussion peut être une source de confusion en l'absence de signaux appropriés. Cependant, la notion de sujet de conversation est débattue en linguistique et n'est pas suffisamment abordée en modélisation du dialogue. Une description précise de cette notion ainsi que de celle de changement de sujet permettrait de comprendre ce que nous percevons lorsque nous jugeons comme cohérente une séquence d'interventions. Afin d'analyser les différents types de changements de sujet, nous proposons de créer un corpus de conversations ayant un but (discussion du dilemme éthique de la *montgolfière*), où les conversations sont des échanges de messages écrits. Ces discussions limitées thématiquement vont nous permettre de comprendre les changements de sujet dans les dialogues.

ABSTRACT

Topics play an important role in dialogue coherence, as what is currently discussed constrains the possible contributions of the participants, and initiating a topic while the previous one is still under discussion may be confusing without appropriate signals. However, how to actually define the notion of topic is debated in linguistics and not sufficiently discussed in dialogue modelling. A precise description of topics and topic shifts in conversation would contribute to understanding what it is we perceive when we judge a sequence of utterances to be coherent. In order to analyse different types of topic shifts, we propose to create a corpus of written task-oriented conversations (discussion of the ethical dilemma of the *balloon task*), where the dialogues happen by message exchanges. Such a controlled setting where the main topic is fixed, and subtopics are more easily identifiable, could be very helpful when it comes to understanding how people change the topic and react to topic shifts in dialogues.

MOTS-CLÉS : Modélisation de dialogue, changements de sujet, sémantique.

KEYWORDS: Topic modelling, Topic shifts, Semantics.

1 Motivation and context

Dialogue is at the core of human interactions, but it is also the place for many misunderstandings. It is thus interesting to build models to represent dialogue and try to better understand the features of a coherent interaction.

Interaction is however complex to characterise and model. In conversations, the participants exchange on different topics and the understanding of the current state of the dialogue by each participant relies on their interpretation of the topic under discussion. Even though topics are usually quite stable, shifts can happen more or less abruptly for several reasons such as a more pressing matter to discuss or the current topic being exhausted. Identifying topic shifts is crucial to understand and take part in the conversation. Conversely, creating confusion by changing the topic untimely and/or in an unclear manner could be seen as uncooperative (Grice, 1975). Moreover, the topic can sometimes gradually change throughout the conversation which can make it complicated to keep track of the most relevant information.

Having a better understanding of topic shifts in dialogue would help us include topics and topic shifts in current dialogue modelling theories (Amblard *et al.*, 2011; Hunter *et al.*, 2015; Maraev *et al.*, 2018; Breitholtz, 2020). It could also enable us to analyse different types of dialogue-like conversations such as message threads on social media and identify off-topic contributions. Moreover, a symbolic approach of dialogue modelling could participate in improving language models' handling of abrupt or unconventional topic shifts.

Topic shifts mechanisms are not all straightforward to investigate as participants usually *implicitly* agree on the topic and negotiate its boundaries through interaction. Thus, a more controlled setting with a fixed main topic, and more easily identifiable subtopics, could be very helpful to understand how people change the topic and react to topic shifts in dialogues.

Topic shifts mostly occur for three reasons: the dialogue content gives an idea to a participant (*e.g.*, remembering an anecdote related to the current topic), the context changes and/or provides something to speak about (*e.g.*, someone is coming who should not hear the current conversation, noises interrupt the dialogue, ...), or a participant has a specific agenda and shifts the topic, potentially abruptly, to fulfil it. In all these cases, the other participants must understand that the topic is changing to prevent disruptions in the dialogue. Specific tasks given to the participants could help us gather different types of topic shifts in our corpus.

Our objective is to construct a corpus that encompasses standard interactions, disrupted interactions, and interactions with multiple goals. The dialogues would be collected with a Chat Tool (Healey & Mills, 2009) that enables the experimenter to intervene in the conversations by modifying or delaying the participants' messages for example. This will facilitate the comparison of how participants handle topic shifts. Our goal is to gather data first in English and later in French and possibly Swedish. Based on these analyses, we will define a proximity measure for determining topic change. Using it, we could segment the dialogues into thematic units and propose a more general theory of interaction.

2 ChatTool

DiET (Dialogue Experiment Toolkit) is an experiment platform developed by Healey & Mills (2009) that provides ways to gather text-messages conversations in a controlled environment. This ChatTool allows experimenter to link participants either with the Telegram app or with a custom computer interface and to store their conversation. All the messages sent by the participants are passed to a server which relays the messages to the addressees. The transition through the server enables the experimenter to control the conversation for they have the possibility to interfere with the conversations. Different manipulations are possible, and both the Telegram app and the custom computer interface have specificities when it comes to the display functionalities.

The experimenter can modify the content of messages, the identity of the sender of the messages and the timing of the messages. They can also insert fake messages and make it look like one of the participants sent them. These manipulations, illustrated by Figure 1, are meant to make it possible to investigate different dialogue phenomenon such as the organisation of a conversation, miscommunication or dialogue coherence.

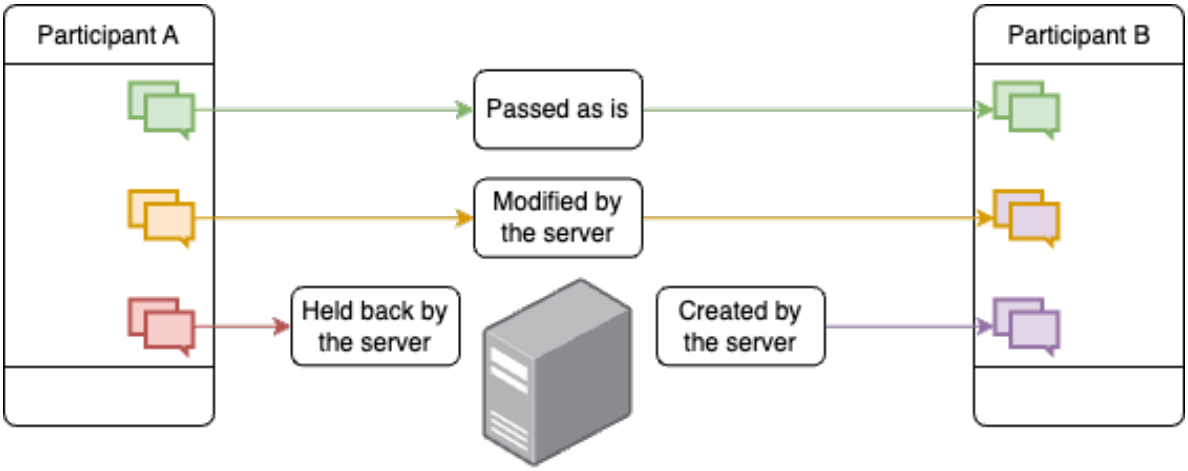


Figure 1: Possible manipulations of the messages with the ChatTool

3 Phenomena to Observe

Our goal is to investigate speakers' understanding of topic shifts. In this context, we would like to analyse the following phenomena.

Topic shifts are not always completely clear, especially when the new topic is not obviously different from the previous one. On the other hand, introducing a new topic that is completely different from the one at hand can be confusing and/or considered rude by the other participants. We would like to see how people react to more or less striking topic shifts.

There exist different ways to mark a topic shift and in particular some linguistic markers such as *by the way* can indicate to the addressees that the speaker is changing the topic. Removing these markers would enable us to assess their influence on mutual understanding.

A topic can change gradually in a conversation or more abruptly. In the first case, each intervention slightly changes the topic and thus helps understanding what is discussed. In the second case, the last interventions can help one notice that the previous topic has been wrapped up which licenses moving to a new one. We would like to see how people cope with missing some parts of the conversation, for instance by preventing one of the participants to receive some of the messages.

The meaning of some contributions can completely change if a piece of them is altered. For example, if someone is trying to describe an image with a cat but every time they use the word 'cat' it is replaced by 'dog', the final message will be completely different. The addressee would picture a different scene in their head and have a different perception of the at-hand topic from what the speaker expected. A message could also become complicated to understand if it contains an argument where one of the premise has been modified. If one receives the message "I need to buy gloves because it is getting so *warm* outside" instead of "I need to buy gloves because it is getting so *cold* outside", their interpretation can be unpredictable. They could believe the sender was being ironic, that they made a mistake, or potentially have an interpretation according to which buying gloves when it gets warm makes sense. We would like to observe the way people interpret messages where a core information has been modified, and in particular if they realise that they have a different understanding of their current discussion or if both speakers try to accommodate the other messages by coming up with a fitting interpretation.

In parallel to these phenomena, we would be interested in understanding how, after how long and with what motivation people shift the topic. This can only be done if the participants create topic shifts on their own as pushing them to change the topic would create a different phenomenon where one has specific reasons to change the current topic.

4 Tasks

This project aims to collect dialogues where participants perform a task involving various subtopics and to analyse the topic shifts and the different phenomena we discussed above.

One such task could be to discuss the ethical dilemma of the *balloon task* (Breitholtz *et al.*, 2021) where participants are asked to choose which of the four passengers of a hot air balloon to throw away, given that they would crash and all die if no one is sacrificed. The subtopics can there be defined by the different arguments in favour or against a passenger. We selected this task because its focus provides a manageable scope for conversation. . We could create more or less obvious topic shifts by inserting a fake message discussing one

of the passengers when another one was discussed, or a message with a different argument from the one currently discussed. We could also create disruption by changing the passenger discussed in certain messages without notification.

A second task where the participants are supposed to cooperate to reach a common goal but where one of them could have hidden intentions would create interactions as both speakers would try to convince the other of their good intentions. Such a task could be realised through a game where the players both own a set of cards representing the wires of a bomb they are trying to disarm. The players know their own cards but not the ones of the other player. Disarming the bomb means cutting the wires in a certain order and thus returning the cards in a certain order. Discussing which card to return every turn would represent a topic. Inserting messages on a specific card or changing the colours that are discussed could create the phenomena discussed in Section 3.

Our last task is a description task inspired from the PhotoBook task (Haber *et al.*, 2019). In this task both participants receive a set of images and they have to determine which one of them they have in common. Describing one image could be considered a sub-topic here. Again, this task provides a manageable scope and should limit the amount of personal information shared by the participants. The modifications of the messages could consist in swapping one entity for another, and we could create more or less clear topic shifts by adding messages related to a different image from the one currently discussed.

All three tasks are designed to prevent the participants from sharing personal information and engaging on personal topics (Amblard *et al.*, 2014). Our goal is to observe the different phenomena described in Section 3 while minimising the impact of our manipulations on the content of the interventions.

5 Experimental Setup

Our goal is to gather conversations for the three tasks described in Section 4 where we manipulate some of the messages to observe the phenomenon discussed in Section 3. Table 1 describes the different manipulations we imagined for all the tasks and phenomena.

In total we have three tasks for four phenomena to observe. We also need one control setting for each task which makes fifteen different experiments. Our goal is to collect about 300 conversations in total (about 20 per experiment setting) where there would be three groups:

- about 30 conversations with first-time participants;
- about 90 conversations with participants joining in for the second, third or fourth time;
- about 180 conversations with 'trained' participants who have already taken part in the experiments more than 4 times.

Phenomenon	Ethical Dilemma Task	Negotiation Task	Image Description Task
Reaction to more or less clear TS	Add message on another passenger / with another argument	Add message on another card	Add message on a very different image from the one discussed / on a similar image
Influence of TS markers	Gather a list of markers and delete them from messages		
Missing information	Delete some arguments	Delete some arguments	Delete some messages with a '?'
Modified messages	Change the passenger discussed	Change the colour of the discussed wire	Modify some entities
Control	No outside interventions		

Table 1: Manipulations of the conversations in the different tasks

We would then annotate the dialogues with the topics but also the type of topic shifts. A topic shift can be more or less brutal and one of our goal is to define a scale where the most abrupt changes can be annotated with e.g. "explicitly marked" and "due to external factors". We also want to annotate the Questions Under Discussion (QUDs) (Kuppevelt, 1995; Ginzburg *et al.*, 1996) in order to compare the granularity that topics and QUDs encode.

6 Conclusion

This paper describes a corpus collection experiment aiming at creating a resource to analyse how people understand each other in more chaotic environments. Our assumption is that speakers are able to maintain interaction in spite of partial or distorted information.

We would like to gather data in three languages: English, French, and Swedish. On top of the contribution in terms of language resources, this would be beneficial to compare how people handle topic shifts in different languages.

The analyses of this corpus would enable us to identify topic shift mechanisms and determine which ones are more easily understandable. This would participate in incorporating topics to formal dialogue models but also in improving language models when it comes to understanding less obvious topic shifts. Our goal in the future is to create a proximity measure to identify topic shifts, which could for instance help find off-topic and potentially harmful contributions in dialogue-like interactions such as social media ones.

References

AMBLARD M., FORT K., MUSIOL M. & REBUSCHI M. (2014). L'impossibilité de l'anonymat dans le cadre de l'analyse du discours. In *Journée ATALA éthique et TAL*, Paris, France. HAL : hal-01079308.

AMBLARD M., MUSIOL M. & REBUSCHI M. (2011). Une analyse basée sur la S-DRT pour la modélisation de dialogues pathologiques. In M. LAFOURCADE & V. PRINCE, Éds., *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles - TALN 2011*, p. 93 – 98, Montpellier, France: Laboratoire d'Informatique de Robotique et de Microélectronique. HAL : hal-00601622.

BREITHOLTZ E. (2020). *Enthymemes and Topoi in Dialogue: The Use of Common Sense Reasoning in Conversation*. Leiden, The Netherlands: Brill. DOI : https://doi.org/10.1163/9789004436794.

BREITHOLTZ E., COOPER R., HOWES C. & LAVELLE M. (2021). *Reasoning in Multiparty Dialogue Involving Patients with Schizophrenia*, In *(In)coherence of Discourse: Formal and Conceptual Issues of Language*, p. 43–63. Springer International Publishing: Cham. DOI : 10.1007/978-3-030-71434-5_3.

GINZBURG J. *et al.* (1996). Dynamics and the semantics of dialogue. *Logic, language and computation*, **1**, 221–237.

GRICE H. P. (1975). Logic and conversation. In P. COLE & J. L. MORGAN, Éds., *Syntax and Semantics*, volume 3, p. 45–47: New York: Academic Press.

HABER J., BAUMGÄRTNER T., TAKMAZ E., GELDERLOOS L., BRUNI E. & FERNÁNDEZ R. (2019). The PhotoBook dataset: Building common ground through visually-grounded dialogue. In A. KORHONEN, D. TRAUM & L. MÀRQUEZ, Éds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 1895–1910, Florence, Italy: Association for Computational Linguistics. DOI : 10.18653/v1/P19-1184.

HEALEY P. G. & MILLS G. J. (2009). A dialogue experimentation toolkit. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 31.

HUNTER J., ASHER N., KOW E., PERRET J. & AFANTENOS S. (2015). Defining the Right Frontier in Multi-Party Dialogue. In *19th Workshop on the semantics and pragmatics of dialogue (SemDial 2015 - goDIAL)*, p. pp. 95–103, Göteborg, Sweden. HAL : hal-01535950.

KUPPEVELT J. V. (1995). Discourse structure, topicality and questioning. *Journal of Linguistics*, **31**(1), 109–147. DOI : 10.1017/S002222670000058X.

MARAEV V., GINZBURG J., LARSSON S., TIAN Y. & BERNARDY J.-P. (2018). Towards kos/ttr-based proof-theoretic dialogue management. In *Proceedings of the 22nd Workshop on the Semantics and Pragmatics of Dialogue*, Aix-en-Provence, France: SEMDIAL.

Actes des journées LIFT 2023
Nancy, France, 20-21 novembre 2023

Décrire et organiser les données de terrain avec RDF

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

RÉSUMÉ

Les données collectées sur le terrain par le linguiste (artefact, enregistrement, specimen, données écrites, etc.) sont diverses et complexes. Il n'existe pas de méthodologie permettant leur description et leur contextualisation. Cette description permettrait aux linguistes de mieux les utiliser ainsi que de les archiver – seule une petite proportion des données de terrain (des paires enregistrements / transcription) peuvent être aujourd'hui archivées. RDF est particulièrement adapté à une description de ces données, notamment dans des ontologies event-centric élaborées par des collectifs de musées ou d'institution d'archive comme *Record in context* ou le CIDOC-CRM. Cette contribution présente les enjeux de cette démarche, les avantages et les inconvénients de cette solution, ainsi qu'une expérience de conversion et d'utilisation de données de terrain en RDF.

ABSTRACT

Linguistic fieldwork data is a complex net of relationships -- for instance, several documents (recordings and theirs related transcriptions) can relate to the same stimuli, several data sessions (each producing written documents, on various supports) can relate to the same recording, or to successives part of it… No methodology or tool is available for the description of this data in a uniform and explicit way. This contribution shows how RDF and an event-centric approach could be used to improve the organization of data by researchers and, then, by archiving bodies. The contextualization of data can offer better interpretability and discoverability of archived content.

MOTS-CLÉS : Linguistique de terrain, données de terrain, métadonnées, RDF, archivage.
KEYWORDS: Fieldwork linguistics, fieldwork data, metadata, RDF, archive.

1 Complexité des données de terrain

Cette contribution porte sur la modélisation et la description des données de terrain linguistique. Ces données sont diverses et complexes : elles comprennent notamment (Evans & Sasse 2003 : 6) :

- des enregistrements (audio, vidéo - souvent en parallèle)

- des données textuelles (cahiers, fichiers) : transcription, listes de mots, notes

- des données géographiques (relevés GPS, carte, toponymes)

- des photographies

- des données bibliographiques

- des spécimens (pierre, roche, insecte, etc.)

- des stimuli

- des dessins

Un inventaire de ces données reste à faire. Mais outre leur variété de types, ces données sont surtout complexes du fait des relations qu'elles entretiennent. Par exemple, plusieurs data sessions (chacune produisant du matériel écrit) peuvent être consacrées à l'analyse d'un même événement langagier (ou à différentes portions du même événement) ; une data session peut être elle-même enregistrée et être interprétée en tant qu'événement communicatif ; plusieurs enregistrements (avec leurs transcriptions éventuelles) peuvent être consacrés au même stimulus. Woodbury (2007) par exemple décrit une situation fréquente chez les linguistes de terrain, ou un texte est enregistré et transcrit plusieurs fois sur une très longue période.

Il n'existe pas d'outil permettant de décrire ces données de façon satisfaisante. L'outil le plus approchant, Lameta (ex-SayMore), développé à l'origine par le Summer Institute of Linguistics, permet de regrouper un enregistrement et une (ou plusieurs) transcription(s) et de leur associer des méta données et des locuteurs. Mais il n'est pas possible d'établir des liens entre les documents, entre leur diverses formes (numérisés, etc), d'associer des transcription à des portions d'enregistrement, de décrire les autres types de données. Les données ne peuvent faire l'objet de requête.

Les chercheurs utilisent essentiellement des solutions *ad hoc*, tel que des tableurs, pour lister les enregistrements collectées et leur associer des propriétés d'une façon rudimentaire. Exposant de bonnes pratiques de gestion des données de terrain, Lee (2022 : 290) montre par exemple l'utilisation d'un tableur ; une colonne « links » est utilisée pour lier les enregistrements à d'autres objets au moyen de relations telle que HasTranscript, Requires, HasPart, isCopy, etc. Le modèle de données (le tableur) est clairement peu approprié pour ce type de données en réseau. Dans mon expérience de linguiste de terrain, tout comme dans les témoignages obtenus auprès d'autres collègues, des solutions de ce type sont utilisées, ne permettant de décrire qu'une portion des données de façon non systématique. L'essentiel des données restent non décrites et sommeille dans des cartons.

Modéliser correctement les données de terrain est pourtant important pour deux raisons au moins. D'une part, il s'agit de pouvoir exploiter correctement et optimalement le matériau collecté. Pouvoir facilement retrouver toutes les productions d'un même locuteur, ou toutes les données produites par des locuteurs de telle tranche d'âge ; ou issus d'un même village, ou de tel type d'événement communicatif, etc. Une meilleure solution de description et de contextualisation des données permettrait d'accroître la « cherchabilité » (*discoverability*) et l'interprétabilité des données.

D'autre part, ces données de terrain sont amenées à être, dans un grand nombre de cas, les seuls témoignages restant de langues bientôt disparues. Dans ce contexte, conserver l'intégralité des données collectées et permettre de les exploiter au mieux philologiquement est un enjeu important. Les sites d'archives de langues du monde (comme par exemple Pangloss, Elar ou Paradisec) n'archivent qu'un petit fragment des données effectivement produites par les linguistes de terrain : des paires enregistrement/transcription. Il s'agit en réalité, plutôt que d'archives, d'un travail d'édition. Pour pouvoir archiver davantage de données, il faut naturellement que les chercheurs produisent la description des données en premier lieu.

Enfin, il y a déjà de nombreux cas de ré-utilisation aujourd'hui d'archives laissées par des chercheurs et produites il y a plusieurs décennies. Là aussi, pour le chercheur qui se plonge aujourd'hui dans les cartons laissés par ses collègues et essaye d'organiser les données, il est nécessaire de disposer de méthodes pour organiser les données – objets, textes, fichiers – que l'on tente d'interpréter au mieux.

2 RDF et les ontologies « event-centric »

RDF semble être un modèle de données bien adapté à la description de ce type de données.

1/ d'une part, des ontologies RDF ont été élaborées par des consortium de musées et d'archives pour décrire et contextualisé des artefacts, par exemple :

- Record in context (https://www.ica.org/standards/RiC/RiC-O_v0-2.html)

- CIDOC-CRM (https://cidoc-crm.org)

Ces ontologies fournissent l'essentielle des relations nécessaires pour décrire ces données. En particulier, elles adoptent une approche « event-centric » : c'est-à-dire que les matériaux produits (enregistrements, transcriptions) sont rattachés à des événements qui les ont produit. Cette approche s'oppose à l'approche « document-centric » où les matériaux produits sont directement reliés entre eux. Le premier modèle est plus lourd mais également plus exhaustif et adapté aux relations philologiquement complexes des données de terrain. Certaines relations ne sont pas proposées par ces ontologies (la relation entre un événement et un artefact « produit par » cet événement, comme des notes de transcription par exemple), et doivent être définies.

2/ D'autre part, RDF permet facilement d'agréger des données issues de sources différentes. C'est le cas des données de terrain, qui seront toujours en partie décrite avec des outils différents (bases de données biographiques, gestions de cartes, etc., de gestion de collection de photo, etc.). RDF permet de récupérer ces données et de les agréger dans une base de connaissance.

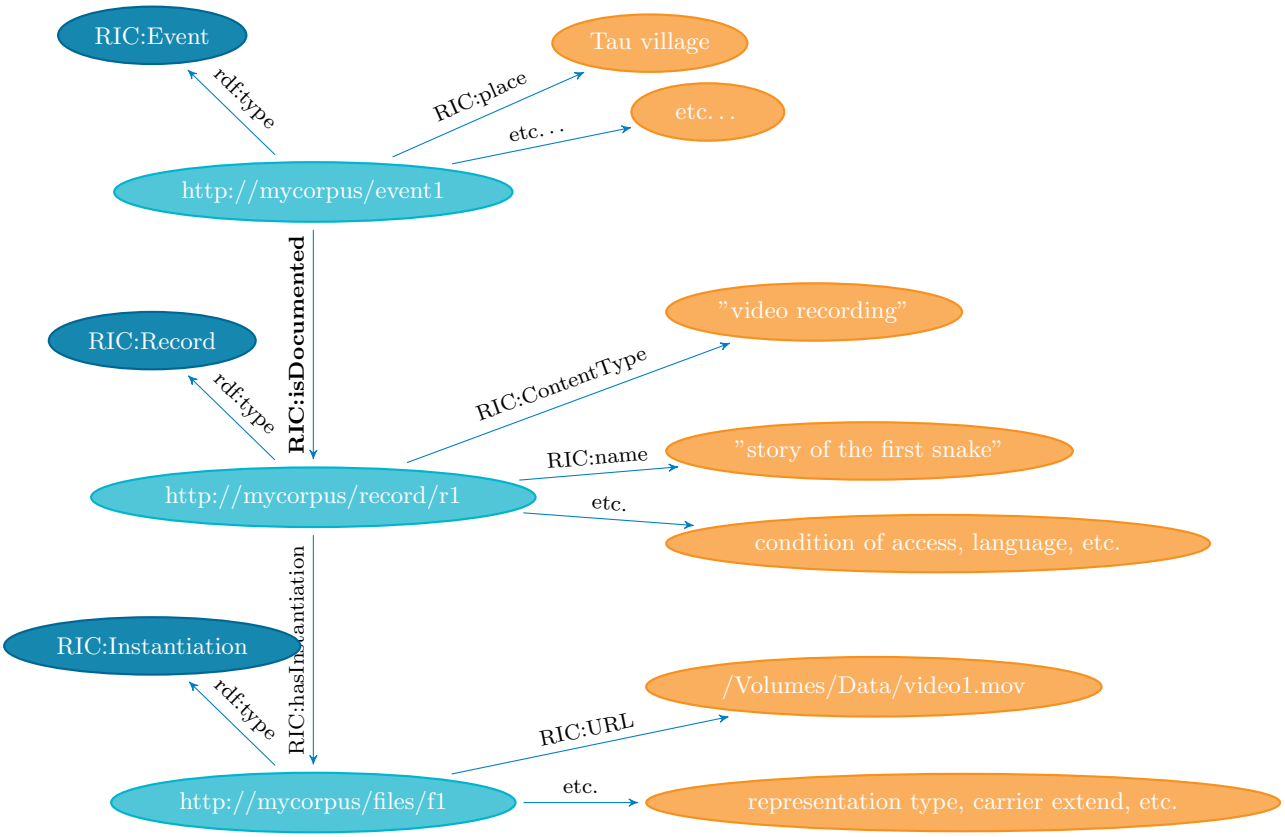
Cette contribution présente une expérience[1] de conversion en RDF et d'enrichissement de données annotées avec différents outils (tableur pour les métadonnées des sessions, Tropy

[1] https://github.com/sylvainloiseau/rdf4lfd

pour les photos[2], Gramps[3] pour les données biographiques, Lameta/SayMore[4] pour les métadonnées des fichiers). Cette expérimentation montre la viabilité de l'utilisation du vocabulaire RDF RIC pour la description des données, et comment de nouvelles relations, spécifiques aux données linguistiques, peuvent être proposées. Elle montre également comment les données ainsi créées peuvent être interrogées avec sparql pour définir des ensembles de données relatifs à un événement, à un locuteur, ou à un stimulus par exemple.

3 Exemple de modélisation avec *Record in context* (RIC)

Nous présentons quelques exemples de modélisation de données de terrain utilisant *Record in context* (RIC). De nombreuses propriétés des objets (date, lieu des événements, sous-typage des événements ou des documents, droits attachés aux documents, etc.) ne sont pas représentés pour ne pas alourdir les graphiques. Un simple événement communicatif ayant fait l'objet d'une captation vidéo peut être représenté ainsi :



Il y a une apparente lourdeur de ce dispositif, mais il permet grâce à sa granularité de représenter les situations complexes, réelles. La distinction distinction *RIC:Event* / *RIC:Record* / *RIC:Instantiation* est utile pour organiser :

[2] https://tropy.org
[3] https://gramps-project.org
[4] https://www.lameta.org

- Des notes manuscrites et leur numérisation (deux *Instantiation* du même *Record*, liées avec une relation *derivedInstantiation*)

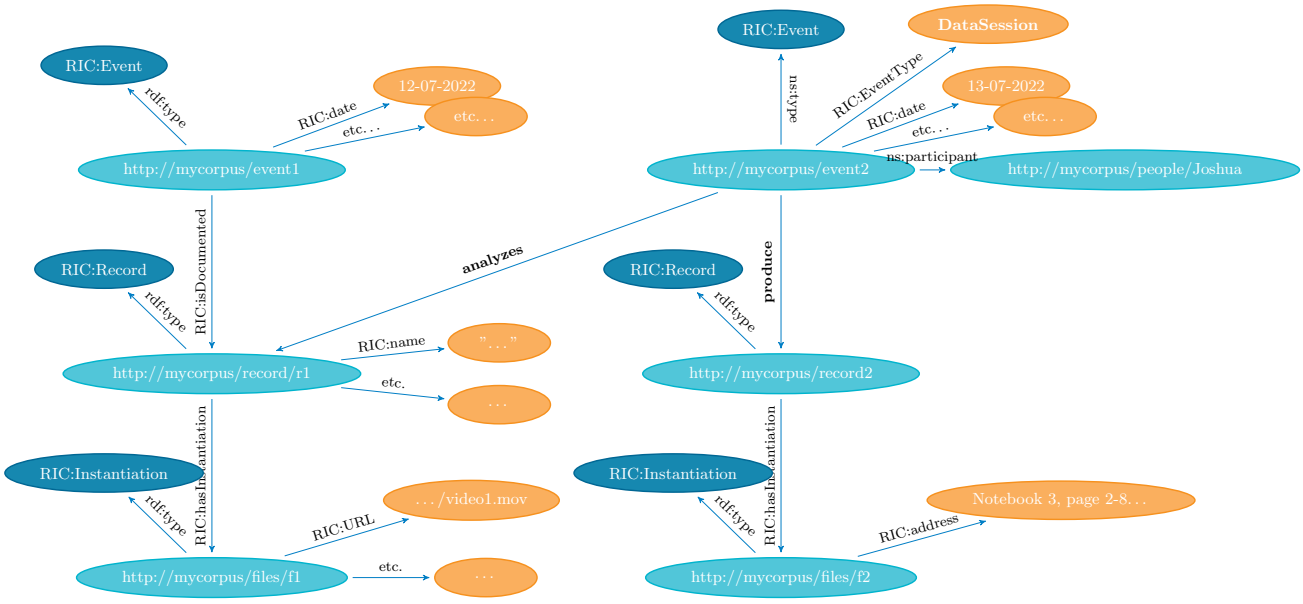
Des transcodages de fichiers multimédia (idem)

- Des prises vidéo successives d'un même événement (plusieurs *Record* reliés à un *Event* et liées entre elles par une relation *successivePart*)

- Des conversion de textes dans différentes logiciels d'annotation (Elan, Flex…)

Illustrons (figure ci-dessous) le cas d'une session de travail consacrée à l'analyse d'un enregistrement. Durant cette session de travail, des notes sont prises dans un cahier (pages 2-8 du cahier 3.) Nous modélisons cela au moyens de deux événements (*RIC:Event*) : l'événement communicatif d'une part et la data session d'autre part (l'attribut *RIC:EventType* permet de typer et distinguer ces événements). Deux nouvelles relations sont introduites : *analyzes*, qui relie une data session et le(s) enregistrement(s) auquel elle est consacrée, et une relation *produces* qui relie l'événement data session et les documents produits durant celle-ci.

Un document (RIC:*Record*) peut donc être lié à un événement (*RIC:Event*) soit avec une relation *RIC:document*, s'il l'enregistre, ou avec une relation *produce*, si le document est produit durant cet événement sans qu'il en soit une captation.



Le modèle s'adapte aisément à toutes sortes de cas de figure, comme par exemple :

- Plusieurs data session consacrées à l'analyse d'une même captation
- Un texte écrit spontanément et donné par un locuteur, puis analysé et lu ("performance" enregistrée) par un autre locuteur. Ici le document écrit pourra être relié à un Événement (*Event*) (où *RIC:EventType=WrittingEvent*) par une relation *Produces*. Des événements de

type (*EventType*) « data session » ou « performance » pourront être reliées au document écrit par des relations *analyses* et *stimulus*.
- Liens vers des stimuli (permettant de grouper les événements liés à un même questionnaire, les différentes versions d'un mythe, etc.) (un Document (*Record*) peut être un stimulus d'un autre *Event*...)
- l'enregistrement d'une data session elle-même
- etc.

RDF rend facile l'agrégation de données issues de différentes sources. Nous avons ainsi agrégé des données issues de Lameta/Saymore, d'une feuille de calcul, d'un système de fichier. Une solution réaliste pour l'utilisation de RDF devra sans doute reposer sur l'utilisation de logiciels d'annotation existant, comme Lameta ou encore Troppy pour les photo, Gramps pour les information biographiques, etc.

Enfin, la base de données peut être facilement interrogée par Sparql pour permettre d'exploiter pleinement les données collectées. Des contextes peuvent collecter de proches en proche tous les nœuds pertinents pour l'interprétation d'un document par exemple (tous les nœuds qui lui sont connectés, directement ou directement, par un chemin Sparql constitué des prédicat (stimulus|documents|analyzes|produces).

Nous avons dû enrichir l'ontologie *Record in Context* de relations reflétant les réalités des archives de terrain linguistiques : les relations *Analyzes*, *Produces*, *Stimuli* (stimulus n'est pas une propriété d'un document (*RIC:Record*), mais une relation entre un document et un événement : tout document (photo, …) peut être utilisé comme stimulus.

La démarche exposée ici peut être rapprochée d'initiatives similaires, comme par exemple le logicisme dans le domaine de l'archéologie, conçu pour la description des sites de fouille (http://www.thearkeotekjournal.org). Une autre approche est RO-Crate, https://www.researchobject.org/ro-crate/1.1/structure.html, approche explorée par l'archive linguistique Paradisec (https://arkisto-platform.github.io/case-studies/paradisec/). L'utilisation de RDF nous semble offrir des avantages décisifs en termes de capacité d'agrégation de données et de requête grâce au langage Sparql, sans parler de l'infrastructure logicielle existante pour ce modèle.

4 Conclusion

En conclusion, un dispositif centré sur RDF pour la description des données de terrain offre des avantages pour les chercheurs comme pour les outils d'archivage des données. Pour les chercheurs, l'enjeux est de rendre réellement utilisable un large pan des données collectées, hétérogènes et fragmentés, difficile à relier et à exploiter pleinement. Pour les archives, l'enjeux est de conserver un ensemble beaucoup plus complet des données des chercheurs, qui permettrait de contextualiser et rendre plus interprétables les données. Les archives se focalisent aujourd'hui sur les paires enregistrement / transcription (ce sont davantage des objets éditoriaux que archivistiques). Si les données archivées doivent permettre des analyses futures, comme suggéré par le paradigme de la documentation linguistique (Himmelmann 1998}), alors il est nécessaire d'archiver davantage de contexte. Des archives futures seront sans doute les seules données disponibles sur des langues menacées. RDF (et Sparql) augment la capacité à chercher (« searchability ») et à explorer

les données (« discoverability ») et permettent d'augmenter l'inter-opérabilité entre archives.

Enfin, notons que pour l'instant un simple inventaire des données produites sur le terrain et de leur relation reste à faire, ainsi qu'une enquête sur les pratiques des chercheurs pour la gestion de leurs données.

Références

Evans, N. and Sasse, H.-J. (2003). Searching for meaning in the library of babel: field semantics and problems of digital archiving. In Barwick, L., Marett, A., Simpson, J., and Harris, A., editors, *Researchers, Communities, Institutions, Sound Recordings*. University of Sydney, Sydney.

Himmelmann, Nikolaus P. (1998). Documentary and descriptive linguistics. In : *Linguistics* 36.1, p. 161-195.

Lee, N. H. (2022). Managing data for writing a reference grammar. In Berez-Kroeker, A. L., Mc-Donnell, B., Koller, E., and Collister, L. B., editors, *The Open Handbook of Linguistic Data Management*, chapter 23, pages 287–299. The MIT Press.

Woodbury, A. C. (2007). On thick translation in linguistic documentation. *Language Documentation and Description*, 4, Peter K. Austin (ed.):120–135.

Actes des journées LIFT 2023
Nancy, France, 20-21 novembre 2023

Décrire une scène ou informer d'un évènement en langue des signes française : des représentations formelles différentes ?

Camille Challant*, Emmanuella Martinod*, Michael Filhol
Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique,
91400, Orsay, France
prénom.nom@lisn.upsaclay.fr

RÉSUMÉ

AZee est un modèle permettant de représenter formellement des discours en langue des signes française. Nous avons décrit avec AZee deux corpus de genres discursifs différents : *40 brèves*, un corpus journalistique, et *Mocap1*, un corpus de description d'images. Traditionnellement considérés en linguistique comme structurellement très différents, les représentations formelles respectives de ces deux corpus présentent-elles des similarités ? Nous montrons que le même ensemble de règles AZee est utilisé dans les deux cas, simplement dans des proportions différentes.

ABSTRACT

Describing a scene or reporting an event in French Sign Language : different formal representations ?

AZee is a model used for formally representing discourses in French Sign Language. We have described with AZee two corpora of different genres : *40 brèves*, a journalistic corpus, and *Mocap1*, a corpus of image descriptions. Traditionally viewed in linguistics as structurally very different, do the respective formal representations of these two corpora show any similarities ? We show that the same set of AZee rules are used in both cases, although with different proportions.

MOTS-CLÉS : Langue des signes française, Représentation formelle, AZee, Modélisation.

KEYWORDS: French Sign Language, Formal representation, AZee, Modeling.

1 Introduction

Notre proposition concerne la modélisation de la langue des signes française (LSF), langue peu dotée et ayant recours à la modalité visuo-gestuelle exclusivement. Comme toute langue des signes (LS), il s'agit d'une langue à part entière se distinguant des langues vocales

*. Les deux auteures ont contribué de façon égale à ce résumé, l'ordre est alphabétique.

Signe TAPIS

Structure illustrative d'un tapis rectangulaire

FIGURE 1

telles que le français par différentes caractéristiques, parmi lesquelles la multilinéarité [1], l'utilisation d'une syntaxe basée sur l'utilisation pertinente de l'espace, ou encore le recours possible à l'iconicité. Deux grands types de structures ont été observés dans *toutes* les LS étudiées à ce jour : les signes lexicaux (i.e. des unités gestuelles ayant un sens générique et conventionnel, que l'on peut produire hors contexte), d'une part, et les structures illustratives, de nature davantage géométrique et iconique, d'autre part. Ces dernières font néanmoins partie intégrante de la langue (Zwitserlood, 2012), malgré la diversité terminologique utilisée pour les décrire (Schembri, 2003; Cormier *et al.*, 2015; Jantunen *et al.*, 2021). De précédents travaux font état de l'utilisation de ces structures dans des proportions variées, selon le genre discursif dans lequel on se trouve (Ferrara & Johnston, 2014; Sallandre, 2014; Jantunen, 2017). Les deux types de structures sont illustrés en figure 1 avec l'exemple d'un *tapis*, pouvant apparaître dans le discours sous la forme d'un signe lexical conventionnel ou d'une structure illustrative qui en montre la forme.

En termes de représentation formelle et de génération des LS, les approches actuelles sont majoritairement inspirées de celles développées pour les langues vocales : ce sont des représentations linéaires basées sur des gloses, ou qui, dans le meilleur des cas, peinent à prendre en compte la complexité possible de la multilinéarité (voir Hadjadj *et al.* (2018) pour un aperçu des approches existantes et de leurs limites). L'approche AZee, dans laquelle s'insère notre travail, vise quant à elle la représentation de l'*ensemble* des phénomènes observables en LS, quel que soit le type de structures concerné.

Dans cette contribution, nous tenterons de comprendre dans quelle mesure les potentielles différences structurelles et de genres discursifs variés se reflètent dans la représentation formelle AZee. Tout d'abord, nous présentons ci-après les grands principes de l'approche AZee et du langage formel qu'elle propose, ainsi que les deux corpus de LSF de genres discursifs différents que nous avons décrits dans le cadre de cette approche (section 2). Ensuite, nous en venons aux résultats du processus de représentation formelle de ces données, à leur comparaison, puis aux pistes de travail à venir (section 3).

1. C'est-à-dire l'utilisation, simultanée ou non, de différents articulateurs manuels et non manuels (par ex. le regard).

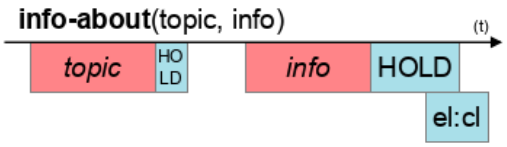
2 Approche théorique et données

AZee (Filhol, 2021) est une approche de description formelle des LS, fondée sur la notion de *règle de production* : une règle qui associe à un sens identifié un ensemble de formes observables à produire. Une règle peut avoir un ou plusieurs arguments. Les règles de production d'une même LS (identifiées grâce à une méthodologie spécifique dans des corpus de LS) forment *un ensemble de production AZee*. Les règles de cet ensemble peuvent se combiner pour construire des *expressions AZee de discours*, qui rendent compte à la fois du sens que l'on interprète et des formes observées d'un discours en LS. Ceci permet de représenter des discours de n'importe quelle taille.

Par exemple, en LSF, la règle de production `chat` associe le sens 'chat' à la forme illustrée en figure 2a. De la même façon, la règle de production `info-about` à deux arguments (*topic* et *info*), associe le sens 'info, à propos de topic' à la synchronisation de formes présentée dans le diagramme 2b [2].

L'expression AZee en figure 2c correspond à la représentation formelle d'une production en LSF dont les formes correspondent à celles de la figure 2b. Elle mobilise les règles `chat` et `dormir`, respectivement en *topic* et *info*, et signifie 'le/un chat dort'.

(a) forme pour 'chat' en LSF



(b) synchronisation de formes pour la règle `info-about`

```
:info-about
    'topic
    :chat
    'info
    :dormir
```

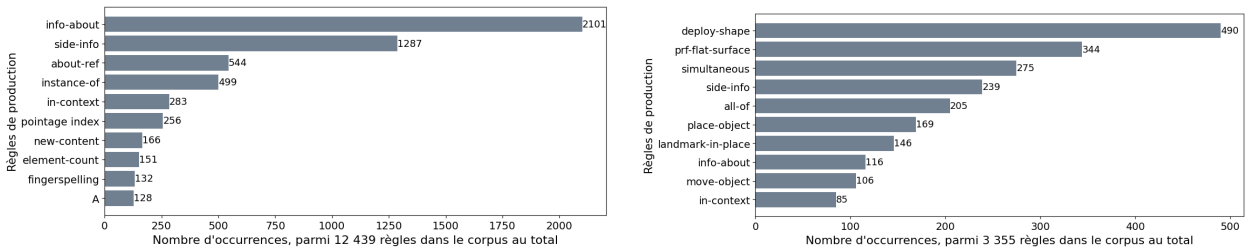
(c) 'le/un chat dort'

FIGURE 2

Nous avons décrit deux corpus avec AZee, correspondant à deux genres discursifs distincts : informatif/journalistique pour l'un, descriptif pour l'autre. Le corpus *40 brèves* (Filhol & Tannier, 2014) est un corpus parallèle français-LSF : 40 brèves journalistiques en français écrit ont chacune été traduites par 3 traducteurs sourds professionnels. Concernant le corpus *Mocap1* (Benchiheub *et al.*, 2016), nous nous sommes focalisés sur une tâche de description de 25 photographies, réalisée par huit locuteurs sourds. Ce travail est encore en cours mais permet d'ores et déjà d'obtenir des résultats substantiels. Ces deux corpus sont chacun représentatif des deux types de structures habituellement identifiés dans les discours : structures essentiellement lexicales pour *40 brèves* et essentiellement illustratives pour *Mocap1*.

2. *HOLD* indique le maintien des arguments pendant une durée déterminée ; el :cl (*eyelid :closed*), un clignement des yeux.

3 Résultats



(a) 40 brèves (b) Mocap1

FIGURE 3 – Comparaison des 10 règles les plus fréquentes dans les deux corpus

Sur la Figure 3, nous pouvons remarquer que les règles de production [3] `info-about`, `side-info` ou encore `in-context` font partie des dix règles les plus utilisées de chaque corpus. Le sens de chacune de ces règles est donné ci-dessous :

— donner une information à propos d'un élément précédemment produit (`info-about`),
— donner une information supplémentaire *mais non essentielle* à propos d'un élément précédemment produit (`side-info`),
— donner un contexte dans lequel se situe la suite de la production (`in-context`).

Malgré les différences de genre discursif, on comprend que ces trois opérations sémantiques peuvent être utilisées indifféremment dans le registre informatif comme descriptif. Ainsi, la même règle `info-about` peut être utilisée dans *40 brèves*, si le locuteur veut préciser qu'un homme est français et, dans *Mocap1*, que la partie basse d'une armoire comporte une porte coulissante.

Néanmoins, certaines règles les plus fréquentes dans *40 brèves* ne font pas partie des plus fréquentes pour la description de *Mocap1* (`about-ref`, `pointage index`, `new-content` [4], etc.) et inversement. Par ailleurs, les règles les plus utilisées dans *Mocap1* (`deploy-shape`, `prf-flat-surface`, `simultaneous`, `place-object`, etc.) sont typiques de la description de scènes : elles mobilisent des arguments de type géométrique tels que *point*, *vecteur*, *trajectoire*, etc. En particulier, `deploy-shape` permet, comme son nom l'indique, de déployer une forme dans l'espace de signation. Cette règle est donc idéale dans le cas de descriptions géométriques de formes spécifiques (par exemple, la forme d'une table particulièrement allongée ou celle d'un lampadaire ancien). `prf-flat-surface` renvoie à une forme de main plate, forme générique utilisée dès lors que l'on veut décrire une surface plate telle qu'un sol ou un mur. On retrouve ces règles dans *40 brèves*, mais beaucoup moins fréquemment. `all-of` est aussi très récurrente dans Mocap1. Cette règle sert en effet à créer une liste d'items, en mettant le focus sur

3. Les règles de production évoquées dans cette partie sont répertoriées dans Challant & Filhol (2022), accompagnées de leur sens.
4. Anciennement, `prise-de-parole`.

l'ensemble en tant que tout. Elle est donc très utile lorsque le signeur veut faire la liste de différents éléments composant une scène ou un paysage. En revanche, une règle comme `fingerspelling` est quasiment absente de la représentation de *Mocap1*. Cette règle servant à l'épellation s'avère nécessaire dans le cas de brèves journalistiques à même de contenir des noms propres peu connus du grand public et qui nécessitent donc d'être épelés. Dans le cadre d'une description de scène, il semble moins probable qu'elle soit utilisée.

De plus, les représentations formelles des différents corpus nous permettent de confirmer les différences de proportions de signes lexicaux au sein de ceux-ci. Dans le cadre de l'approche AZee, nous considérons comme lexicale toute règle de production sans argument obligatoire. Ces règles (e.g. `chat`, cf. section 2) peuvent se réaliser sans faire appel à des éléments extérieurs, contrairement aux règles comportant des arguments obligatoires (e.g. `info-about`). Dans *40 brèves*, 43% des règles de production correspondent à cette définition (sur l'ensemble des règles utilisées dans le corpus), contre 10% seulement dans *Mocap1*. Les signes lexicaux sont donc présents en moindre mesure dans ce second corpus.

Enfin, nous notons également que le corpus *40 brèves*, pour lequel on s'attendait à retrouver une majorité d'éléments lexicaux, comporte finalement seulement 43% de règles de production lexicales. Les 57% de règles de production restantes correspondent à des règles telles que `info-about` ou `in-context`. Ces règles permettent de générer simultanément de nombreuses informations, que ce soit au niveau de la forme (clignements des yeux, inclinaison de la tête, maintien de la position des mains, etc.), mais aussi du sens ('ajout d'une information', 'déploiement d'une forme', etc.). Une annotation plus traditionnelle aurait nécessité l'ajout de gloses spécifiquement dédiées à chacune de ces informations, là où le système AZee semble plus économique. Ce dernier constat souligne le fait que des données de LSF a priori aisément annotables à l'aide de gloses constituent finalement des phénomènes linguistiques plus complexes qu'il n'y paraît.

4 Conclusion et perspectives

Nos résultats soulignent les potentialités d'AZee pour décrire des données de LSF comportant pourtant des structures linguistiques très différentes et rarement formalisées ailleurs autrement que par des gloses. Les mêmes règles de production sont utilisées pour représenter les énoncés provenant des deux corpus distincts, bien que ces règles soient mobilisées dans des proportions différentes. Ceci corrobore certaines observations de linguistes des LS évoquant un continuum entre le lexique et les structures illustratives des LS [5]. Les structures illustratives semblent en effet régies par des principes organisationnels comparables à ceux du lexique. AZee met cela en avant et permet de mieux catégoriser et quantifier ces phénomènes. Dans la continuité de Filhol & McDonald (2020), notre travail, à la jonction entre linguistique et informatique, contribue à confirmer par une méthologie différente ces

5. Voir entre autres Cuxac (1985); Johnston & Schembri (1999); Boutet *et al.* (2010); Zwitserlood (2012).

précédentes observations.

De futurs travaux pourraient mettre en avant le fait que l'observation de proportions différentes de règles utilisées constitue un indice du type de registre dans lequel on se trouve. Ainsi, nous prévoyons de finaliser l'écriture des expressions AZee pour *Mocap1*, puis d'effectuer le même travail pour d'autres genres discursifs (narratif, explicatif, etc.), afin de comparer l'utilisation des différentes règles pour chaque genre, en suivant la même méthodologie. Pour finir, l'annotation de corpus avec AZee est également pertinente dans le domaine de la synthèse à l'aide de signeurs virtuels, comme le montrent notamment les travaux de Sharma (2023); Sharma & Filhol (2023).

Références

BENCHIHEUB M.-E.-F., BERRET B. & BRAFFORT A. (2016). Collecting and Analysing a Motion-capture Corpus of French Sign Language. In *sign-lang@ LREC 2016*, p. 7–12 : European Language Resources Association (ELRA).

BOUTET D., SALLANDRE M.-A. & FUSELLIER-SOUZA I. (2010). Gestualité humaine et langues des signes : entre continuum et variations. **131**, 55–74.

CHALLANT C. & FILHOL M. (2022). A First Corpus of AZee Discourse Expressions. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, p. 1560-1565, Marseille, France.

CORMIER K., SMITH S. & SEVCIKOVA SEHYR Z. (2015). Rethinking Constructed Action. *Sign Language Linguistics*, **18**(2), 167–204.

CUXAC C. (1985). Esquisse d'une typologie des langues des signes. volume Autour de la langue des signes, Journées d'Études 10. UFR de linguistique générale et appliquée., p. 35–60, Université René Descartes, Paris.

FERRARA L. & JOHNSTON T. (2014). Elaborating who's what : A Study of Sonstructed Action and Clause Structure in Auslan (Australian Sign Language. *Australian journal of linguistics*, **34**(2), 193–215.

FILHOL M. (2021). *Modélisation, traitement automatique et outillage logiciel des langues des signes*. Habilitation à diriger des recherches, Université Paris-Saclay. HAL : tel-03197108.

FILHOL M. & MCDONALD J. (2020). The Synthesis of Complex Shape Deployments in Sign Language. In *Proceedings of the 9th workshop on the Representation and Processing of Sign Languages*.

FILHOL M. & TANNIER X. (2014). Construction of a French–Lsf Corpus. In *Building and Using Comparable Corpora, Language Resource and Evaluation Conference (LREC)*, p. 2–5, Reykjavik, Iceland.

HADJADJ M., FILHOL M. & BRAFFORT A. (2018). Modeling French Sign Language : a Proposal for a Semantically Compositional System. In *International Conference on Language Resources and Evaluation*.

JANTUNEN T. (2017). Constructed Action, the Clause and the Nature of Syntax in Finnish Sign Language. *Open Linguistics*, **3**.

JANTUNEN T., DE WEERDT D., BURGER B. & PUUPPONEN A. (2021). The more you move, the more action you construct : A Motion Capture Study on Head and Upper-torso Movements in Constructed Action in Finnish Sign Language narratives. *Gesture*, **19**(1), 76–101.

JOHNSTON T. & SCHEMBRI A. (1999). On Defining Lexeme in a Signed Language. *Sign language linguistics*, **2**(2), 115–185.

SALLANDRE M.-A. (2014). *Compositionnalité des unités sémantiques en langues des signes. Perspective typologique et développementale*. Thèse hdr, Université Paris 8.

SCHEMBRI A. (2003). *Rethinking 'classifiers' in signed languages*, In *Perspectives on classifier constructions in sign languages*, p. 3–34. K. D. Emmorey : Mahwah, NJ : Lawrence Erlbaum Associates.

SHARMA P. (2023). A Layered Approach to Constrain Signing Avatars. In *VISIGRAPP_DC 2023*, Lisbon, Portugal : Scitevents. HAL : hal-04143663.

SHARMA P. & FILHOL M. (2023). Intermediate block generation for multi-track sign language synthesis. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, p. 1–2.

ZWITSERLOOD I. (2012). *Classifiers*, In *Sign language : An international handbook*, p. 158–186. Mouton de Gruyter.

Actes des journées LIFT 2023
Nancy, France, 20-21 novembre 2023

Est-ce que l'extraction des interrogatives du français peut-elle être automatisée ?

Valentin D. Richard[1]
(1) LORIA, Université de Lorraine, 615 Rue du Jardin-Botanique, 54506 Vandœuvre-lès-Nancy, France
valentin.richard@loria.fr

RÉSUMÉ

La quasi totalité des études linguistiques sur les interrogatives du français se contente d'extraire ces dernières d'un corpus à la main ou grâce à de simples heuristiques basées sur du texte brut (mots interrogatifs, point d'interrogation,...). Dans ce papier, je présente FUDIA (French UD Interrogative Annotator), un programme qui permet de détecter les interrogatives du français d'un corpus annoté en dépendances universelles (UD). FUDIA est un système de réécriture de graphe par règles, basé sur Grew. Je liste les obstacles à une telle tâche d'identification automatique des interrogatives et j'explique comment FUDIA en résout la plupart. Je montre que, couplé à un parseur affiné sur des données similaires, FUDIA obtient de bons résultats sur du texte brut (écrit et transcription de l'oral).

ABSTRACT

Can French Interrogative Retrieval be Fully Machine-Based ?

The vast majority of linguistic corpus studies on French interrogatives retrieve the researched patterns by hand or only based on simple heuristics on raw text (e.g. interrogative words, question marks). In this paper, I present FUDIA (French UD Interrogative Annotator), a program able to detect French interrogatives from a corpus annotated in Universal Dependencies (UD). FUDIA is a rule-based graph rewriting system based on Grew. I inventory the obstacles to such an interrogative identification task and I explain how FUDIA solves most of them. I show that, coupled with a parser fine-tuned on similar data, FUDIA obtains good results on raw text (written and speech transcription).

MOTS-CLÉS : interrogatives, français, Universal Dependencies, par règles, réécriture de graphe.

KEYWORDS: interrogatives, French, Universal Dependencies, rule-based, graph rewriting.

1 Introduction

Parmi les études de corpus récentes portant sur les interrogatives du français, la plupart récupèrent leurs données en les extrayant à la main (Reinhardt, 2019a; Bally, 2022). Quelques

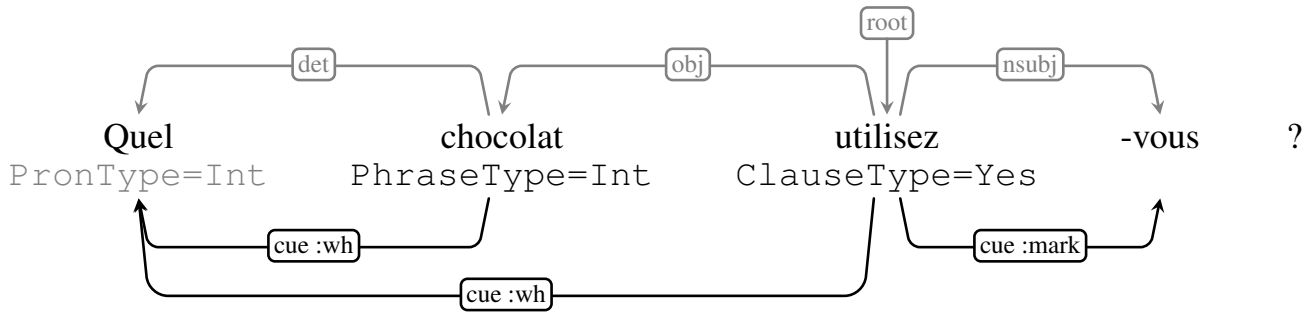


FIGURE 1 – Annotation d'une question de GSD par FUDIA. Les arcs et traits d'origine sont en gris. Les arcs `cue` peuvent être retirés pour conserver la structure d'arbre initiale.

méthodes semi-automatiques sont employées, comme l'utilisation d'un concordancier (Rossi-Gensane *et al.*, 2021; Benzitoun, 2022; Gillet, 2022), d'expressions régulières (Lefeuvre & Rossi-Gensane, 2017) ou d'heuristiques simples (Reinhardt, 2019b; Eshkol-Taravella *et al.*, 2022) (mots interrogatifs et points d'interrogations). Seule Lefeuvre (2021) profite d'un corpus (LM10 (Habert, 2005) [1]) préalablement annoté syntaxiquement grâce à l'outil Syntex (Bourigault *et al.*, 2005) pour identifier certains motifs.

2 Proposition

FUDIA Le programme FUDIA (French UD Interrogative Annotator) s'appuie sur la présence de corpus déjà annotés en syntaxe, notamment les corpus francophones du projet Universal Dependencies (UD) (de Marneffe *et al.*, 2021) : FQB (Seddah & Candito, 2016), GSD (Guillaume *et al.*, 2019), ParisStories (Kahane *et al.*, 2021), ParTUT (Bosco & Sanguinetti, 2014), PUD (McDonald *et al.*, 2013), Rhapsodie (Lacheret *et al.*, 2014) et Sequoia (Candito & Seddah, 2012). FUDIA contient des règles de réécriture de graphe basées sur Grew [2] (Bonfante *et al.*, 2018). Ces règles listent les structures attestées d'interrogatives en français [3].

Annotations et FIB À partir de la représentation UD d'une phrase, FUDIA rajoute le trait `ClauseType=Int` indiquant sur la tête d'une proposition (finie, infinitive ou averbale) qu'elle est interrogative. Les mots interrogatifs et marquages morphosyntaxiques (inversion du clitique sujet, *est-ce que* / *si*,...) sont aussi identifiés par des dépendances spécifiques, appelée arcs `cue` (voir Fig. 1).

Les interrogatives des corpus francophones UD ont été extraites en un corpus appelé French

1. Voir aussi `http://redac.univ-tlse2.fr/voisinsdelemonde/index.jsp`
2. `https://grew.fr/`
3. On s'intéresse ici aux interrogatives d'un point de vue syntaxique, comme définies dans la Grande Grammaire du Français (GGF) (Delaveau *et al.*, 2021).

Interrogative Bank (FIB). À l'aide d'un script fourni, il est possible d'y calculer automatiquement la proportion d'un certain type d'interrogative, selon la classification de Coveney (2011). Le FIB étend le FQB (French Question Bank) en y apportant une plus grande diversité de structures syntaxiques (ex. des phrase de la forme : *Est-ce que* + sujet + verbe) et des interrogatives enchâssées.

Le code source de FUDIA [4] ainsi que le French Interrogative Bank [5] sont librement disponible en ligne.

3 Difficultés

Obstacles Les difficultés rencontrées pour correctement identifier les interrogatives sont nombreuses. On recense des obstacles linguistiques, notamment la grande variabilité de ces constructions en français, et les formes proches en apparence (ex. inversion stylistique pour rapporter des paroles). Du fait de son format, UD a quelques limites. Par exemple, rien ne différencie lexicalement les *si* interrogatifs des *si* conditionnels. Mais un bon nombre de difficultés est attribuable à la variabilité et la non-uniformité des annotations UD. Par exemple, on a retrouvé 6 annotations différentes de *est-ce que* sur 36 occurrences, dont aucune avec la relation `fixed`, censée être employée pour les expression figées.

Solutions apportées Le développement de FUDIA a cherché à prendre en compte toutes les formes attestées d'interrogatives dans la littérature scientifique, mêmes celles non standards, comme le titre de ce papier (*est-ce que* + inversion du clitique). De plus, le programme évite le plus possible de dépendre de listes de mots de classe ouverte. Notamment il ne présume pas des verbes pouvant enchâsser une interrogative. Le faire risquerait de louper des occurrences non envisagées. Cette stratégie s'est avérée payante car elle a permis de "découvrir" 6 verbes introducteurs d'interrogative listés ni dans la GGF (Delaveau *et al.*, 2021, Tab. XXII-7 p.1414) ni par Defrancq (2005, chap. 1 n.b.p. 11). Ces verbes sont : *connaitre*, *enseigner*, *interroger*, *mesurer*, *souligner* et *tester*.

Les obstacles linguistiques et les limitations d'UD sont en partie résolues grâce à quelques heuristiques sur l'environnement syntaxique des structures recherchées, par exemple la forme du verbe principal ou la distinction complément / ajout. La présence de nombreux cas et exceptions a pu être traitée grâce à un schéma de disjonction de motifs écrit en python. Quelques expressions clés (ex. *est-ce que* et *qu'est-ce que*) sont réannotées de manière uniforme en suivant les théories de la GGF (Delaveau *et al.*, 2021) et les choix d'annotation du corpus CEFC/Orféo [6] (Benzitoun *et al.*, 2016).

4. https://github.com/Valentin-D-Richard/FUDIA
5. https://github.com/Valentin-D-Richard/UD_French-FIB
6. https://repository.ortolang.fr/api/content/cefc-orfeo/11/documentation/site-orfeo/guide-dannotation-syntaxique-du-corpus-orfeo/

	exactitude	précision	rappel	F1
FUDIA	0,905	0,966	0,770	0,857
sur l'écrit seul	0,860	0,917	0,647	0,759
sur l'oral seul	0,950	1,00	0,875	0,933

TABLE 1 – Score de FUDIA sur la détection d'interrogatives du français

4 Évaluation

Méthode Pour évaluer FUDIA, un corpus de 200 phrases a été constitué. Il contient des phrases tirées aléatoirement de corpus écrit (Annodis (Péry-Woodley *et al.*, 2011)), oral (OFROM (Avanzi *et al.*, 2012)), de questions (Maya (Reinhardt, 2016), TenNovels (Reinhardt, 2019b)) et d'interrogatives enchâssées (Defrancq, 2005). Une tâche d'annotation regroupant 12 participant·es les a annoté selon si elles contiennent au moins une interrogative ou pas d'interrogative du tout. On utilise ces étiquettes de référence pour calculer le score de FUDIA.

En premier lieu, les phrases sont annotées en UD par un parseur pré-entrainé grâce à ArboratorGrew [7]. La moitié du corpus d'évaluation issu de corpus écrits est parsée en affinant le parseur avec GSD (1476 phrases, LAS = 0,922). L'autre moitié, issue de corpus oraux, en affinant sur Rhapsodie et ParisStories (total : 2675 phrases, LAS = 0,818). Puis FUDIA est exécuté sur la sortie.

Résultats La tâche d'annotation obtient un bon score inter-annotateur·rice : Cohen κ minimum = 0,613, moyenne = 0,781, maximum = 0,924. Cependant, quelques phrases ont généré plus de désaccord. On compte 11 phrases sur 200 avec un écart-type élevé (supérieur à 0,47, c.à.d. au moins un tiers de désaccord).

Les résultats de l'évaluation de FUDIA, sur l'ensemble du corpus d'évaluation ainsi que sur chacune des parties écrites et orales de celui-ci, sont affichés en Table 1.

5 Interprétation

Tâche d'annotation Parmi les phrases engendrant le plus de désaccord, on trouve certaines interrogatives enchâssées, ex. (1-a) (crochets rajoutés pour délimiter l'interrogative). Ces structures apparaissent donc comme plus difficiles à détecter pour les humains. L'autre type majeur de débat concerne les marqueurs du discours insistant sur l'attente d'une réponse, tels

index.html

7. https://arboratorgrew.elizia.net/

hein ? ou *non ?* (1-b). Les consignes demandaient d'étiqueter les déclaratives questionnantes, qui auraient une intonation montante à l'oral, comme non-interrogatives. Mais le statut de ces marqueurs n'y était pas suffisamment précisé.

(1) a. À travers un récit largement autobiographique, le comique français a décrit avec humour [comment nombre des « gauchistes » d'alors ont troqué le manteau afghan et les sabots hollandais pour la veste et l'attaché-case d'aujourd'hui]. (Defrancq écrit)
b. Tu me fais confiance, non ? (TenNovels)

Scores de FUDIA La précision de FUDIA est élevée. Nous attribuons ça aux nombreuses heuristiques qui permettent d'éliminer les structures qui ressemblent à des interrogatives mais n'en sont pas.

Le point faible de FUDIA est son rappel. En tout, 17 interrogatives ne sont pas correctement identifiées (12 de la partie écrite du corpus dévaluation, 5 de la partie orale). Parmi ces faux négatifs, on estime que la quasi totalité sont dus au parseur en amont. L'erreur la plus fréquente concerne les mots QU annotés comme des mots relatifs (ex. (2-a)) ou des conjonctions de subordination, au lieu de comme des mots interrogatifs. Par exemple, dans (2-b) , le parseur annote les deux *comment* en tant que conjonction de subordination. C'est erroné et surprenant, car *comment* est toujours un adverbe en français, et toutes les occurrences de *comment* dans les données d'affinage sont annotées comme des adverbes.

(2) a. mais il fallait répondre à qui (Defrancq oral)
b. Nous ne voulons pas savoir comment les faits se sont déroulés, mais comment ils auraient pu ou dû arriver. (Defrancq écrit)
c. [...] l'AIEA n'était pas en mesure de vérifier s'il y a eu détournement ou non de matériel nucléaire [...] (Defrancq écrit)
d. ce qui me préoccupe c'est que va-t-on faire avec XXX (Defrancq oral)

Un autre type d'erreur concerne la confusion entre circonstancielle conditionnelle et interrogative complément. Par exemple, dans (2-c), le parseur a annoté le syntagme "*s'il y a eu détournement ou non de matériel nucléaire*" comme une proposition ajout au lieu d'une proposition complément de *vérifier*. De ce fait, FUDIA la classifie comme une conditionnelle, sans détecter la séquence *ou pas*.

Le dernier type de faux négatif est lié à la tokénisation. J'ai utilisé `spaCy` pour tokéniser le corpus d'évaluation. Mais le traitement du *-t-* euphonique dans l'inversion sujet-verbe n'y est pas le même que dans les données d'affinage. Ainsi, dans la séquence tokénisé *va -t -on* de (2-d), le parseur a annoté *-t* comme sujet explétif de *va* et *-on* comme son objet indirect. FUDIA ne détecte donc pas d'inversion avec un sujet appartenant à la liste des pronoms

clitiques.

Enfin, il est aussi étonnant à première vue de noter que FUDIA performe moins bien sur l'écrit que sur l'oral, et ce d'autant plus que le score LAS [8] du parseur affiné sur de l'écrit était supérieur à celui affiné sur de l'oral. Au vu des phrases et des types d'erreurs discutés ci-dessus, cela s'explique surement par la plus grande complexité en moyennes des phrases du sous-corpus écrit (longueur, structures enchâssées, relations à longue distance, etc.).

6 Conclusion

Je répondrais à la question du titre par l'affirmative. Les corpus arborés nous permettent de produire des programmes, comme FUDIA, qui détectent les interrogatives du français avec un bon score. Cependant, la tâche reste difficile, et FUDIA peut être moins performant sur des données bruitées ou très variables (ex. productions d'enfants (Gillet, 2022)). De plus, les performances (dont celles des parseurs) dépendent beaucoup de la qualité des annotations en entrée. C'est pourquoi la pérennité, le maintien et la révision constante des corpus arborés me semblent encore une des tâches essentielles de notre discipline.

Remerciements

Je remercie très fortement Bruno Guillaume, †Guy Perrier et Sylvain Kahane pour leur aide et leurs commentaires sur mon travail.

Références

AVANZI M., BÉGUELIN M.-J., CORMINBOEUF G., DIÉMOZ F. & JOHNSEN L. A. (2012). OFROM – corpus oral de français de Suisse romande.

BALLY A.-S. (2022). Les interrogatives totales en français québécois dans l'écrit SMS : à la croisée de l'oral et de l'écrit. In F. NEVEU, P. PRÉVOST, A. STEUCKARDT, G. BERGOUNIOUX & B. HAMMA, Éds., *8e Congrès Mondial de Linguistique Française*, volume 138, p. 12006, Orléans : SHS Web of Conferences. DOI : 10.1051/shsconf/202213812006.

BENZITOUN C. (2022). Évolution des interrogatives partielles directes en français : le cas de combien. In *8e Congrès Mondial de Linguistique Française*, volume 138 de *Syntaxe*, p. 13003, Orléans : SHS Web of Conferences. DOI : 10.1051/shsconf/202213813003.

BENZITOUN C., DEBAISIEUX J.-M. & DEULOFEU H.-J. (2016). Le projet ORFÉO : un corpus d'étude pour le français contemporain. *Corpus*, **15**(15). DOI : 10.4000/corpus.2936.

8. *labeled attachment score* : pourcentage de tokens ayant la bonne étiquette et le bon gouverneur.

BONFANTE G., GUILLAUME B. & PERRIER G. (2018). *Application of Graph Rewriting to Natural Language Processing*, volume 1. ISTE Wiley.

BOSCO C. & SANGUINETTI M. (2014). Towards a Universal Stanford Dependencies parallel treebank. In V. HENRICH, E. HINRICHS, D. DE KOK, P. OSENOVA & A. PRZEPIÓRKOWSKI, Éds., *Proceedings of the 13th Workshop on Treebanks and Linguistic Theories (TLT-13)*, p. 14–25, Tübingen (Germany).

BOURIGAULT D., FABRE C., FRÉROT C., JACQUES M.-P. & OZDOWSKA S. (2005). Syntex, analyseur syntaxique de corpus. In *Actes Des 12èmes Journées Sur Le Traitement Automatique Des Langues Naturelles*, Dourdan, France : Association pour le Traitement Automatique des Langues.

CANDITO M. & SEDDAH D. (2012). Le corpus Sequoia : Annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Actes de La 19e Conférence Sur Le Traitement Automatique Des Langues Naturelles*, p. 321–334, Grenoble, France : Association pour le Traitement Automatique des Langues.

COVENEY A. (2011). L'interrogation directe. *Travaux de linguistique*, **63**(2), 112–145.

DE MARNEFFE M.-C., MANNING C. D., NIVRE J. & ZEMAN D. (2021). Universal Dependencies. *Computational Linguistics*, **47**(2), 255–308. DOI : 10.1162/coli_a_00402.

DEFRANCQ B. (2005). *L'interrogative enchâssée. Structure et interprétation*. Champs linguistiques. Louvain-la-Neuve : De Boeck Supérieur.

DELAVEAU A., CAPPEAU P. & DAGNAC A. (2021). Les phrases interrogatives. In A. ABEILLÉ & D. GODARD, Éds., *La Grande Grammaire du Français*, volume 2, p. 1402–1437. Arles : Actes Sud/Imprimeries nationales Éditions, 1 édition.

ESHKOL-TARAVELLA I., BARBEDETTE A., LIU X. & SOUMAH V.-G. (2022). Classification automatique de questions spontanées vs. préparées dans des transcriptions de l'oral. In Y. ESTÈVE, T. JIMÉNEZ, T. PARCOLLET & M. ZANON BOITO, Éds., *Traitement Automatique Des Langues Naturelles*, p. 305–314, Avignon, France : ATALA.

GILLET P. (2022). Développement du langage de l'enfant : L'exemple des interrogatives partielles. In *8e Congrès Mondial de Linguistique Française*, volume 138, p. 13002, Orléans : SHS Web of Conferences. DOI : 10.1051/shsconf/202213813002.

GUILLAUME B., DE MARNEFFE M.-C. & PERRIER G. (2019). Conversion et améliorations de corpus du français annotés en Universal Dependencies. *Revue TAL*, **60**(2), 71.

HABERT B. (2005). Text corpus of "Le Monde". DOI : https ://www.islrn.org/resources/421-401-527-366-2/.

KAHANE S., CARON B., STRICKLAND E. & GERDES K. (2021). Annotation guidelines of UD and SUD treebanks for spoken corpora : A proposal. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, p. 35–47, Sofia, Bulgaria : Association for Computational Linguistics.

LACHERET A., KAHANE S., BELIAO J., DISTER A., GERDES K., GOLDMAN J.-P., OBIN N., PIETRANDREA P. & TCHOBANOV A. (2014). Rhapsodie : un Treebank annoté pour l'étude de l'interface syntaxe-prosodie en français parlé. In *4e Congrès Mondial de Linguistique Française*, volume 8, p. 2675–2689, Berlin, Germany : SHS Web of Conferences. DOI : 10.1051/shsconf/20140801305.

LEFEUVRE F. (2021). Les interrogatives averbales dans la presse, stratégies discursives récurrentes ? *Langue francaise*, **212**(4), 107–122.

LEFEUVRE F. & ROSSI-GENSANE N. (2017). Les interrogatives indirectes en discours informel oral. *Langue française*, **196**(4), 51–74. DOI : 10.3917/lf.196.0051.

MCDONALD R., NIVRE J., QUIRMBACH-BRUNDAGE Y., GOLDBERG Y., DAS D., GANCHEV K., HALL K., PETROV S., ZHANG H., TÄCKSTRÖM O., BEDINI C., BERTOMEU CASTELLÓ N. & LEE J. (2013). Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 92–97, Sofia, Bulgaria : Association for Computational Linguistics.

PÉRY-WOODLEY M.-P., AFANTENOS S., HO-DAC L.-M. & ASHER N. (2011). La ressource ANNODIS, un corpus enrichi d'annotations discursives. *Revue TAL*, **52**(3), 71.

REINHARDT J. (2016). Établir un corpus oral de questions : L'analyse semi-automatisée avec Praat et Perl à l'exemple de cinq épisodes de Maya l'Abeille. In *5e Congrès Mondial de Linguistique Française*, volume 27, p. 11007, Tours : SHS Web of Conferences. DOI : 10.1051/shsconf/20162711007.

REINHARDT J. (2019a). La transmission du manque d'information dans la téléréalité française. In *Transmission, oubli et mémoire dans les sciences du langage, JC2017 - 20èmes Rencontres des jeunes chercheurs en Sciences du Langage*, Paris, France.

REINHARDT J. (2019b). Les interrogatives directes tirées de dix romans policier. DOI : https ://hdl.handle.net/11403/interrogatives-in-novels/v1.

ROSSI-GENSANE N., CÓRDOBA L. F. A., URSI B. & LAMBERT M. (2021). Les structures interrogatives directes partielles fondées sur *où* dans les dialogues de romans français du XXe siècle. *Journal of French Language Studies*, **31**(2), 169–191. DOI : 10.1017/S0959269520000253.

SEDDAH D. & CANDITO M. (2016). Hard Time Parsing Questions : Building a QuestionBank for French. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 2366–2370, Portorož, Slovenia : European Language Resources Association (ELRA).

Actes des journées LIFT 2023
Nancy, France, 20-21 novembre 2023

Vers un repérage automatique des discontinuités dans le discours pathologique du sujet schizophrène

Barrouillet Vincent-Thomas[1] Michel Musiol[1, 2] Maxime Amblard[2]
(1) Universite de Lorraine, CNRS, ATILF, F-54000 Nancy, France
(2) Universite de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
vincent-thomas.barrouillet@univ-lorraine.fr

RESUME

Le discours du sujet schizophrène en phase aiguë présente de multiples bizarreries. Un sous-ensemble de ces dernières a été modélisé en tant que « ruptures ». En étude systématique de corpus, on dégage 27 séquences comportant 47 ruptures dans des entretiens entre des patients et des psychologues. Nous définissons un modèle qui vise à retrouver de façon *in fine* automatique ces ruptures. À partir de la segmentation en actes des tours de parole, nous définissons les actes pertinents pour chaque interactant au moyen de règles de décision non ambiguës. Enfin, un calcul algorithmique permet de repérer les ruptures attendues. Nous obtenons une F-mesure de 88.9 % et une exactitude de 96.2 %. L'analyse qualitative des données montre que les 40 vrais positifs et les 3 faux positifs relèvent d'une même catégorie, les « ruptures thématiques », tandis que le système échoue à repérer les autres catégories de ruptures.

ABSTRACT

Toward an automatic identification of discontinuities in the pathological discourse of patient with schizophrenia

Pathological discourse of schizophrenics in acute phase presents multiple oddities. A subset of them was modeled as "discontinuities". Here, 27 sequences with 47 discontinuities were manually identified in our corpus of interviews between a patient and a psychologist. We aim to find these discontinuities *in fine* automatically. From the segmentation of the turns into acts, we precise all the relevant acts for each interactant and each turn, thanks to unambiguous decision rules. Finally, an algorithm identifies the discontinuities. We achieve a F-measure of 88.9% and an accuracy of 96.2%. By analyzing the data, we point out that the 40 true positives and the 3 false positives fall into the same category, "thematic discontinuity", while the system fails to identify all other categories of discontinuities.

MOTS-CLES : discours, discours pathologique, schizophrénie, corpus
KEYWORDS: discourse, pathological discourse, schizophrenia, corpus

1 Contextualisation

Le discours pathologique du sujet schizophrène est parsemé de bizarreries que l'on repère intuitivement. Un sous-ensemble de ces bizarreries a été formalisé sous la forme de ruptures, qui sont des discontinuités pragmatiques dans le discours. Un modèle général a ainsi émergé identifiant quatre types de ruptures (Musiol, 2009) qui sont le débrayage conversationnel et la défectuosité de l'initiative conversationnelle d'une part, et les ruptures inter et intra d'autre part. Les deux premières sont prédites par un modèle adapté de la SDRT (Asher & Lascarides, 2003; Rebuschi et al., 2014). Ici, nous nous intéressons plus particulièrement à la partie du modèle prévoyant les deux dernières catégories : les ruptures inter, i.e. ayant lieu entre deux tours de parole, et les ruptures intra, i.e. ayant lieu à l'intérieur d'un tour de parole. Ces ruptures sont objectivement et manuellement mises à jour au moyen de la modélisation hiérarchique (Roulet et al., 2001), mais nous cherchons à les identifier par un système *in fine* automatique qui ne passe plus par la modélisation hiérarchique.

2 Présentation du corpus

Le corpus est constitué d'entretiens qui se sont tenus à l'hôpital psychiatrique de Rouen en 2005 et à l'hôpital de Tizi-Ouzou, en Algérie, en 2019. Nous avons obtenu un CPP (« Comité de Protection de la Personne ») pour le corpus de Rouen, mais l'absence de comité d'éthique à Tizi Ouzou ne nous a pas permis d'en présenter un. Aussi, afin de garantir la qualité de la méthodologie appliquée, nous avons réemployé un protocole pour lequel nous avions obtenu un accord la même année auprès du CHRU d'Aix-En-Provence. Le protocole consiste en la réalisation d'un ou deux entretiens non directifs, sans but explicite, permettant au sujet d'élaborer sur son environnement et sa vie quotidienne. Les entretiens sont conduits par plusieurs psychologues ayant reçu les formations nécessaires.

Le corpus, qui rassemble 26 heures d'entretiens, a ensuite été retranscrit manuellement par quatre transcripteurs entrainés. Un guide de transcription a été élaboré pour normaliser la production, et les deux transcripteurs ont été entraînés sur des données parallèles afin de contrôler leur production. Nous n'avons pas calculé le Kappa de Cohen car leur entraînement conjoint a été fait sur un volume réduit de données. Le matériel textuel obtenu est désidentifié des noms de personne, de ville, d'institution, etc., et augmenté de balises *ad hoc* telles que <rire> <chuchotement> <claquement de porte>, etc.

Une étude manuelle par deux spécialistes différents des transcripteurs a été effectuée, en application du modèle de (Roulet et al., 2001), pour identifier l'ensemble exhaustif des séquences où les ruptures apparaissent. L'étude a permis d'identifier 27 séquences de ruptures. Les séquences sont composées de plusieurs tours de parole que nous segmentons

en actes de langage (Searle & Vanderveken, 1985). Comme il n'existe pas de système automatique ayant les performances suffisantes, la segmentation est réalisée par un annotateur, et le résultat est corrigé et validé par un second annotateur.

3 Méthodologie

Notre méthode est basée sur la notion de « mise à jour ». Nous assumons qu'un acte est mis à jour, au sens large, par un acte d'un tour ultérieur du même interactant si le second a explicitement besoin du premier pour exister, comme dans un couple question / réponse, ou dans le cas d'une reformulation, d'une reprise, d'une élaboration, ou encore d'une répétition. Nous envisageons l'un après l'autre l'ensemble des actes d'un tour pour l'un puis pour l'autre des interactants et nous formalisons 5 règles explicitant les traitements à appliquer à chaque acte. Si un acte n'est pas mis à jour par la suite, il n'est pas retenu comme acte pertinent (règle A1). S'il est mis à jour par la suite pour le même interactant, il est retenu comme acte pertinent de tour en tour pour cet interactant (règle A3) jusqu'à sa mise à jour où il n'est plus pertinent (règle A4). L'ensemble des autres règles s'applique pour statuer sur la pertinence éventuelle de l'acte qui acte met à jour le précédent. Et il n'y a aucun acte pertinent dans le dernier tour de parole de la séquence (règle A5).

Toutes les règles font l'objet d'une description formelle, mais la règle A2 reste la plus complexe d'un point de vue conceptuel. Chaque tour de parole possède un et un seul acte dit "prévalent". L'acte prévalent est l'acte « le plus important » du tour de parole qui se calcule comme celui dont les autres actes dépendent. Il est toujours possible de retirer soit un, soit un groupe d'actes secondaires, mais la suppression de l'acte prévalent dans le tour de parole lui retire la possibilité de faire sens.

Ce concept est une modélisation de l'« acte principal » défini dans le modèle hiérarchique de (Roulet et al., 2001) et des travaux en cours visent à automatiser le calcul de l'acte prévalent, afin de se libérer totalement de la modélisation hiérarchique dans la détection automatique les ruptures. La règle A2 stipule que l'acte prévalent d'un interlocuteur est forcément pertinent pour le tour en cours et pour l'interlocuteur concerné. Et s'il n'est pas mis à jour par la suite, il n'apparaitra qu'une fois dans le tour de parole qu'il structure.

Nous utilisons un tableau à deux entrées contenant les tours de parole pour les deux interactants. Dans chaque case se trouvent les actes pertinents à échéance d'un tour de parole pour le patient ou le psychologue. Les actes pertinents en vertu de la règle A3 portent en indice le numéro de l'acte qui le mettra à jour, et les actes pertinents en vertu de la règle A2 ne portent pas d'indice, car ils n'apparaissent qu'une fois.

Par la suite, un algorithme détermine parmi les actes pertinents ceux qui ne doivent pas être pris en compte dans le calcul des ruptures, et qu'il faut retirer — ces actes sont appelés « actes fantômes ». Enfin, si les deux ensembles d'actes restants, patient et psychologue, ne

sont pas identiques pour un tour de parole donné, le système repère une rupture à échéance de ce tour.

4 Pseudo-code de l'algorithme de détection des ruptures

Dans cette section, nous proposons une description algorithmique de l'étape finale de la détection des ruptures. Nous commençons par les notations initiales.

R := [R0, R1, ..., Rk] # le tableau complet R se compose des tours de parole R0 à Rk

Ri := [Ri0, Ri1] # le tour Ri se compose des cases du patient (0) et du psychologue (1)

Rij := [A0, A1, ...] # la case Rij se compose d'une liste d'actes A0, A1, …

An := [corps, indice] # l'acte An se définit comme un corps et un indice

corps := étiquette # le corps d'un acte est une étiquette qui en donne le numéro (A, B, C…)

indice := rien ou étiquette # l'indice d'un acte est l'étiquette de l'acte qui le mettra à jour. Il n'y a pas d'étiquette en cas d'application de la règle A2

Nous poursuivons par la définition d'une fonction supprimant les éléments non pertinents.

fonction Rij_sans_fantôme(i, j): # retirer les actes fantômes de R(i, j)

. liste_fantômes est initialisé comme liste vide

. pour tous les actes A de R(i, j)

. . pour tous les actes B de R(i, 1 – j)

. . . si A = B :

. . . . alors : passer au A suivant

. . . . sinon : pour tous les actes C de R(i – 1, 1 – j)

. . . . . si (corps de A = corps de C) et (corps de B = indice de C) alors :

. . . . . A est placé dans liste_fantômes

. pour tous les actes de R(i, j), retirer les fantômes et renvoyer le résultat

fin de la fonction

Enfin, nous abordons la fonction qui identifie les ruptures.

fonction trouver_rupture(k): # k est le nombre de tours

. pour i allant de 0 à k # pour chaque tour i

. . pour les deux valeurs de j # les deux interactants

. . . R(i, j) prend la valeur de Rij_sans_fantome(i, j)

. . . si R(i, 0) différent de R(i, 1), alors, il y a rupture à échéance du tour i

fin de la fonction

5 Exemple : Voyage & Incontinence

Nous développons dans la suite un exemple complet d'analyse d'une rupture.

Actes de langage par tour de parole	Interactant S $R_{i,0}$	Interactant P $R_{i,1}$
S_0 A = J'ai toujours une belle-mère et un beau-père. **B** = Mon beau-père, depuis qu'il est à la retraire, s'absente neuf mois… C = Et quand il revient, il est violent.	$R_{0,0}$ $\{B_M ; C_O\}$	$R_{0,1}$ $\{B_D ; C_D\}$
P_0 **D** = Il va où pendant ces neuf mois ?	$R_{1,0}$ $\{B_M ; C_O ; D_E\}$	$R_{1,1}$ $\{D_F\}$
S_1 **E** = Oh, je ne sais pas où il va, puisque je ne l'ai jamais accompagné…	$R_{2,0}$ $\{B_M ; C_O ; E\}$	$R_{2,1}$ $\{E_F ; D_F\}$
P_1 **F** = Il vous dit pas ?	$R_{3,0}$ $\{B_M ; C_O\}$	$R_{3,1}$ $\{F_L\}$
S_2 **G** = Il est incontinent. H = Hachem. I = Il est incontinent mon beau-père. J = Mon beau-père est incontinent comme un bébé. K = Ou un petit garçon jusqu'à trois ans.	$R_{4,0}$ $\{B_M ; C_O ; G\}$	$R_{4,1}$ $\{F_L\}$
P_2 **L** = Et du coup, il s'absente 9 mois ?	$R_{5,0}$ $\{B_M ; C_O ; L_M\}$	$R_{5,1}$ $\{L\}$
S_3 **M** = Depuis qu'il est à la retraite, oui, il voyage.	$R_{6,0}$	$R_{6,1}$

N = Il s'absente 9 mois O = Et quand il revient, il est violent. … U = Et je suis tombée. V = Et j'ai saigné.	Ø	Ø

Les Ri,j du tableau sont l'ensemble des actes pertinents au regard des règles de décision. L'algorithme donné ci-dessus calcule les actes fantômes qui sont représentés ici en vert. Lorsque Ri,0 et Ri,1 ne sont pas identiques, après le retrait des actes fantômes, la rupture est repérée, ici pour i = 3 et pour i = 4, ce qui signifie qu'il y a une rupture, ici inter, à échéance de ces tours de paroles. Le tour de la patiente noté S2 est donc encadré par deux ruptures thématiques et se constitue donc une sorte d'ilot isolé du reste de l'interaction.

6 Résultats et discussion

Parmi les 27 séquences du corpus, le système repère 40 vrais positifs et 213 vrais négatifs, ainsi que 3 faux positifs et 7 faux négatifs. On obtient donc une exactitude de 96,2 % et une F-mesure de 88,9 %, ce qui est un bon résultat au vu de la complexité de la tâche. Plus finement, le système a une précision de 93.0 % et un rappel un peu inférieur, à 85.1 %.

L'analyse qualitative des résultats conduit à plusieurs remarques. Dans l'interaction, les ruptures peuvent être de nature différente : ruptures thématiques ou argumentatives. L'ensemble des positifs (vrais et faux) sont des ruptures thématiques, contrairement à l'ensemble des faux négatifs. Le système ne reconnait aucune rupture argumentative, et se focalise sur les ruptures thématiques, y compris les 3 fausses. Nous disposons ainsi d'un système formel permettant d'établir la présence de ruptures thématiques avec un très bon niveau de performance. Mais si ces ruptures sont les plus nombreuses dans le corpus, le système nécessite une adaptation plus fine pour identifier les autres types de ruptures qui apparaissent au niveau de l'interaction.

7 Conclusion

Les « ruptures », dans le discours pathologique du sujet schizophrène, sont des points de discontinuité pragmatique que l'on peut mettre au jour objectivement et manuellement au moyen de la modélisation hiérarchique (Roulet et al., 2001). Dans le cadre d'un corpus d'entretiens patients et psychologue, et afin d'automatiser le repérage des ruptures dans le but de se passer de la modélisation hiérarchique, nous proposons un système fondé sur la notion de « mise à jour » : nous assumons qu'un acte de langage est mis à jour par un acte postérieur si le second a besoin du premier pour exister (e.g. un couple question / réponse).

Cinq règles formelles décident pour chaque acte de chaque tour de parole, pour le patient et le psychologue, si l'acte est « pertinent » ou pas. Si c'est le cas, l'acte se retrouve dans une case Ri,j d'un tableau à deux entrée, i étant le numéro du tour et j valant 0 pour le patient, 1 pour le psychologue. Un algorithme, dont nous avons donné le pseudocode, calcule alors les ruptures dans la séquence, et nous obtenons sur les 27 séquences du corpus, comportant 47 ruptures, une exactitude de 96,2 % et une F-mesure de 88,9 %, soit de bons résultats au regard de la complexité de la tâche.

Références

ASHER N. & LASCARIDES A. (2003) Logics of conversation. *Cambridge University Press*.

MUSIOL M. (2009) Chapitre 13. Incohérence et formes psychopathologiques dans l'interaction verbale schizophrénique *in Psychose, langage et action. Louvain-la-Neuve*, pp. 217–238. De Boeck Supérieur (Neurosciences & cognition).

REBUSCHI M., AMBLARD M. & Musiol, M. (2014) Using SDRT to analyze pathological conversations: Logicality, rationality, and pragmatic deviances *in Interdisciplinary works in logic, epistemology, psychology and linguistics*, pp. 343–368. Springer.

Roulet E. et al (2001) Un modèle et un instrument d'analyse de l'organisation du discours, vol 62. Peter Lang Gmbh, collection *« Sciences pour la communication »*. Berne.

SEARLE J.R. & VANDERVEKEN D. (1985) Foundations of illocutionary logic. *CUP Archive*.

Troisième partie

Session dédiée aux ressources et application

Actes des journées LIFT 2023
Nancy, France, 20-21 novembre 2023

A Semi-supervised Dialogue Discourse Parsing Pipeline

Chuyuan Li[1] Maxime Amblard[1] Chloé Braud[2]
(1) LORIA, UMR 7503, Université de Lorraine, CNRS, Inria, 54000 Nancy, France
lisa27chuyuan@gmail.com, maxime.amblard@loria.fr
(2) IRIT, Université de Toulouse, CNRS, ANITI, Toulouse, France
chloe.braud@irit.fr

RÉSUMÉ

Analyse semi-supervisée du discours dans les dialogues

L'analyse du discours joue un rôle crucial dans le Traitement Automatique des Langues (TAL) et a démontré son utilité dans diverses applications telles que le résumé et les systèmes de questions-réponses. Dans cet article, nous abordons ce problème difficile en raison de la rareté des données annotées : l'analyse du discours dans les dialogues. Notre approche de l'analyse du discours comporte deux étapes : tout d'abord, nous prédisons la structure du discours, puis nous identifions les relations au sein de la structure. En utilisant seulement 50 exemples comme données d'entraînement, nos méthodes obtiennent des résultats compétitifs par rapport à l'état de l'art supervisé dans le même domaine et de bien meilleures performances inter-domaines, avec également une meilleure stabilité.

ABSTRACT

Discourse analysis plays a crucial role in Natural Language Processing (NLP) and has demonstrated its usefulness in various downstream applications like summarization and question answering. In this work, we study discourse in dialogues : an under-explored setting due to significant data scarcity challenge. We conduct discourse parsing within a pipeline : first, we predict the discourse structure, and then we identify the relations within the structure. Using only 50 examples as gold training data, our methods achieve competitive results compared to supervised state-of-the-art in-domain and much stronger performance cross-domain, with also better stability.

MOTS-CLÉS : Analyse du discours, apprentissage automatique, dialogue.

KEYWORDS: Discourse analysis, machine learning, dialogue.

1 Introduction

Discourse analysis aims to uncover the inherent structure of documents and has been shown useful for many applications, from sentiment analysis or fake news detection (Bhatia *et al.*, 2015; Karimi & Tang, 2019), to summarization or machine translation (Chen & Yang,

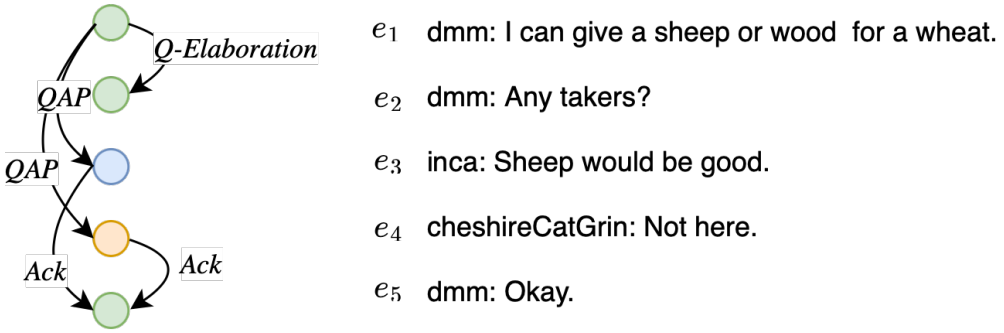


FIGURE 1 – An SDRT graph structure (left) of a dialogue (right) in STAC corpus (*s2-leagueM-game4*). e are EDUs. QAP : *question answer pair* ; Ack : *acknowledgment* ; Q-Elaboration : *question elaboration*. Graphic extracted from Li (2023).

2021; Chen *et al.*, 2020). In recent years, the availability of accurate transcription methods and the increase in online communication have led to a tremendous rise in dialogue data, necessitating the development of automatic analysis systems. However, simple surface-level features are oftentimes not sufficient to extract valuable information from conversations (Qin *et al.*, 2017). Rather, we need to understand the semantic and pragmatic relationships organizing the dialogue, for example through the use of discourse information.

Discourse parsing task consists of retrieving a structure from documents, where spans of text are linked by semantic-pragmatic relations (such as *Explanation, Acknowledgment, Contrast*...). It is a hard task, with low performance especially for multi-party dialogues involving intricate relations between speakers. Hence, on the English chat corpus STAC (Asher *et al.*, 2016) (board game) – annotated under the Segmented Discourse Representation framework (Asher & Lascarides, 2003) with graph structures and 16 relations –, State-Of-The-Art (SOTA) supervised parser Structured-Joint (Chi & Rudnicky, 2022) reports 59.6 at best on the full structure, with a drop of about 20 points for cross-domain when testing on the Molweni corpus (Li *et al.*, 2020) (Ubuntu forum). A main challenge in discourse parsing is data scarcity, along with limitations of supervised approaches in cross-domain scenarios (Liu & Chen, 2021) or incomplete parsing that overlooks the important relation information (Badene *et al.*, 2019; Huber & Carenini, 2022; Li *et al.*, 2023). In this work, we propose the first semi-supervised full discourse parsing pipeline that sequentially conducts parsing tasks. We show that with minimal supervision, our pipeline can achieve comparable results to supervised models both in in-domain and cross-domain scenarios.

2 Preliminaries

2.1 Segmented Discourse Representation Theory

The Segmented Discourse Representation Theory (SDRT) (Asher, 1993; Lascarides & Asher, 1993; Asher & Lascarides, 2003) is a dynamic representation theory of discourse. In SDRT,

	Split			#DU/doc		#Tok/sent		#Tok/doc		#Spk/doc		Rel
Dataset	train	dev	test	max	avg	max	avg	max	avg	max	avg	type
STAC	947	105	109	105	13.0	13	4.4	607	50	6	3.0	16
Molweni	9,000	500	500	14	8.8	17	11.9	208	105	9	3.5	16

TABLE 1 – Statistics in STAC and Molweni corpora. Numbers of discourse units per document (DU/doc), tokens per sentence (tok/sent), tokens per document (tok/doc), speakers per document (spk/doc) are given. Both corpora have the same relation types.

the basic elements of analysis are clause-like text spans, known as Discourse Units (DUs). The smallest units of DUs are Elementary Discourse Units (short in EDUs). The coherence of a document is obtained via a structure – oftentimes tree-like or graph-like – of rhetorically connected discourse units. Figure 1 gives an SDRT-annotated example where nodes and edges represent EDUs and relations, respectively.

2.2 Discourse Datasets

We utilize two English dialogue corpora in this study, both annotated under the SDRT framework. Some key statistics are shown in Table 1.

The Strategic Conversations corpus (STAC) (Asher *et al.*, 2016) is currently the most commonly used dialogue corpus to train SDRT-style parsers. It contains 45 online multi-party strategic chat logs during the board game *The Settlers of Catan*, where players discuss and exchange resources to build roads and cities. The vocabulary in this corpus is thus special, with a high frequency of words such as *sheep, clay, wood*. The corpus is manually annotated and divided into 1161 sub-documents. We follow the split in Shi & Huang (2019) : 947 for training, 105 for validation, and 109 for testing.

The Molweni Corpus (Li *et al.*, 2020) is derived from the Ubuntu Chat Corpus (Lowe *et al.*, 2015), where 10,000 short multi-party technical chat logs are annotated for discourse analysis and machine reading comprehension. Despite its large size, a large portion of the documents are highly repetitive. The original annotation suffers from quality issues such as inconsistency, making the results less reliable. Therefore, we revised the annotation of a small subset (50 documents) to ensure a more robust evaluation (test only).

3 Proposition and Experiments

3.1 A Pipeline Design

A standard discourse parsing involves three tasks : EDU segmentation, link attachment, and relation prediction. Most previous work in dialogue discourse parsing starts with gold-

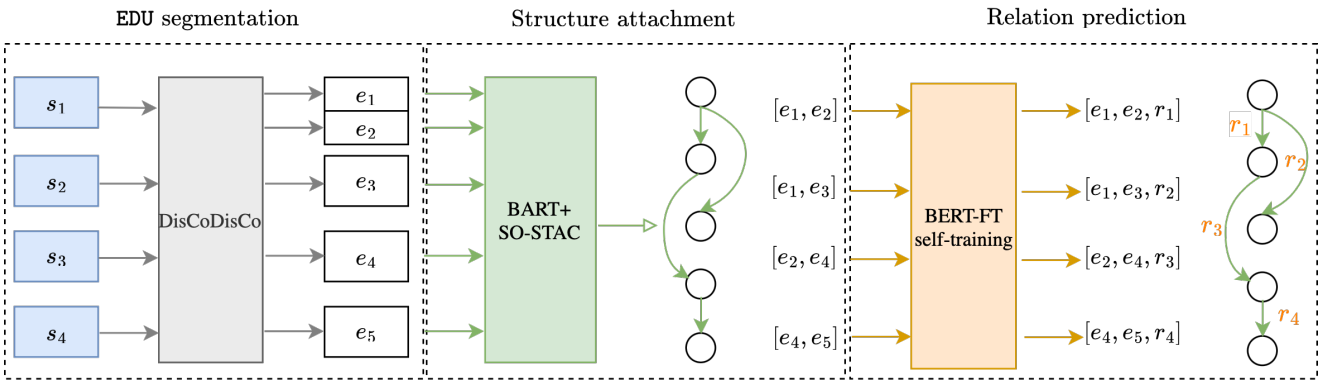


FIGURE 2 – Semi-supervised discourse parsing pipeline proposition. s are utterances ; e are EDUs ; r are rhetorical relations. DisCoDisCo model is proposed in Gessler *et al.* (2021). BART+SO-STAC is BART model fine-tuned on Sentence Ordering task (Li *et al.*, 2023). BERT-FT is BERT model fine-tuned with self-training for relation prediction.

standard EDU annotations and applies a *structure-then-relation* approach (Afantenos *et al.*, 2015; Shi & Huang, 2019; Liu & Chen, 2021; Wang *et al.*, 2021). We follow this pipeline by first predicting naked structures and then providing relations, as depicted in Figure 2. Remarkably, we train the system using only 50 documents in STAC, with an average of 13 EDUs per document, making it the first semi-supervised discourse parsing system for dialogues.

Although most previous work begins with gold EDUs, we consider it crucial to evaluate in a deployed scenario where the parser performs segmentation first. We thus integrate an off-the-shelf segmenter DisCoDisCo (Gessler *et al.*, 2021) – a straightforward sequence tagging model – to perform EDU segmentation. DisCoDisCo achieves an F_1 score of 94.8%.

Next, the predicted EDUs are put into a fine-tuned BART model (Lewis *et al.*, 2020) for **Structure Attachment**. BART is fine-tuned with dialogue-tailored Sentence Ordering task (Barzilay & Lapata, 2008) to enhance the pair-wise, inter-speech block, and inter-speaker discourse information. We hypothesize that the location of discourse information in the network may vary, possibly influenced by the length and complexity of the dialogues. Therefore, we investigate each attention head individually. In one attention head, the attention values among EDUs can be seen as edge weights (Liu & Lapata, 2018). Thus, by using Maximum Spanning Tree algorithms such as Eisner (Eisner, 1996), we obtain a discourse tree structure. For the key issue on choosing the best attention head, we use a small validation set of {10, 30, 50} annotated documents. We find that with just 50 examples, the optimal attention head can be consistently located.

The last step is **Relation Prediction** : with predicted EDU pairs, the goal is to assign a rhetorical relation among 16 candidates. Here we choose BERT model (Devlin *et al.*, 2019) as backbone as it has shown superior performance in discourse-related classification tasks (Chen *et al.*, 2019; Atwell *et al.*, 2021). We prepare the input relation pairs by following the Next Sentence Prediction pattern in BERT, inspired by Shi & Demberg (2019) : a [CLS]

Model	#Train	Segment	Link	Relation	Full
SJ (Chi & Rudnicky, 2022)	1000	-	$70.7_{0.5}$	$77.3_{1.2}$	$54.6_{0.7}$
SJ (Chi & Rudnicky, 2022)	50	-	$55.1_{3.5}$	$61.1_{2.1}$	$33.6_{2.2}$
Ours w gold EDU & link	50	-	-	$58.4_{1.3}$	-
Ours w gold EDU	50	-	$\mathbf{59.3_{0.7}}$	$\mathbf{62.0_{1.1}}$	$\mathbf{38.6_{0.7}}$
Ours w pred EDU	50	94.8	$52.2_{0.4}$	$61.2_{1.6}$	$32.8_{0.9}$

TABLE 2 – Semi-supervised parsing results with the reproduction of SOTA supervised parser Structured Joint (SJ) and our semi-supervised pipeline. Scores are average micro-F_1 over 10 runs. In 50 train setup, best scores are in bold. - not applicable.

token begins the sequence, followed by the first EDU, [SEP], and the second EDU. We loosely translate the output probabilities in BERT model as its predictive confidence, enabling sorting predicted pairs. We select the top k pairs of most confident pseudo-labeled data in each relation type, in which way we maintain the label ratios. This is a simple yet effective sample selection criterion. Through iterative self-training, our classifier is enhanced with the combination of gold and pseudo-labeled data.

3.2 Full Parsing Results

The full parsing results on STAC test set are displayed in Table 2. For comparison, we replicate the SOTA supervised model Structured Joint (SJ) (Chi & Rudnicky, 2022) which uses RoBERTa-base model (Liu *et al.*, 2019) as backbone and employs 3-dimension attention to encode links and relations jointly. In the upper part of the Table, we show SJ performance with 1000 and 50 training data. In the lower part, we detail relation prediction results (gold EDU & link), parsing with gold segmentation (gold EDU), and parsing with predicted segments (pred EDU). Anecdotally, the extracted structures on STAC corpus are found to be similar to the gold SDRT-graphs, achieving an F_1 score of 59.3 and outperforming a strong baseline by 3 points (Li *et al.*, 2023). For relation prediction, our self-trained BERT classifier achieves an accuracy of 58.4% at best. When applying the deployed pipeline, we obtain 32.8% micro-F_1, as displayed in the last line of Table 2. Under the same training size, our pipeline exhibits much better performance compared to SJ model in both link attachment (59.3% vs. 55.1%) and relation prediction (62.0% vs. 61.1%) tasks, bringing a noteworthy improvement of 5 points in full parsing.

In order to test the generalizability of our proposal, we apply SJ model and our pipeline in a cross-domain setup : training on 50 documents from STAC and evaluate on 50 re-annotated dialogues in Molweni test set (Molweni-clean). Preliminary results on Molweni-clean show that our pipeline achieves superior performance on all tasks, surpassing SJ model on link (+24%), relation (+8%), and full parsing (+14%). On relation prediction, SJ considers the

tree structure and relation jointly, while our approach focuses on individual relation pairs. As documents across different genres exhibit diverse structures, our method, despite being more localized, is better suited for general applicability. Moreover, our model exhibits greater stability, whereas the SJ model is heavily biased towards one domain.

4 Conclusion

In this work, we propose a versatile pipeline for sequentially addressing all tasks in discourse parsing. In conformity with real-world situations with limited labeled data, we leverage information from Pre-trained Language Models such as BART and BERT and utilize semi-supervised techniques. Our method shows strong performance in both in-domain and cross-domain settings.

For future work, we intend to improve the derived discourse structures and explore the use of more, possibly out-of-domain raw data, and investigate other bootstrapping approaches for relation prediction. We would also like to evaluate our pipeline on different kinds of data such as transcribed spoken dialogues and another discourse-annotated framework such as the Rhetorical Structure Theory.

Acknowledgements

The authors thank the anonymous reviewers for their insightful comments and suggestions. This work was supported by the PIA project “Lorraine Université d’Excellence”, ANR-15-IDEX-04-LUE, as well as the CPER LCHN (Contrat de Plan État- Région - Langues, Connaissances et Humanités Numériques). It was partially supported by the ANR (ANR-19-PI3A-0004) through the AI Inter-disciplinary Institute, ANITI, as a part of France’s “Investing for the Future — PIA3” program, and through the project AnDiAMO (ANR-21-CE23- 0020). We would like to thank the Grid’5000 community (https ://www.grid5000.fr/).

Références

AFANTENOS S., KOW E., ASHER N. & PERRET J. (2015). Discourse parsing for multi-party chat dialogues. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 928–937, Lisbon, Portugal : Association for Computational Linguistics. DOI : 10.18653/v1/D15-1109.

ASHER N. (1993). *Reference to abstract objects in discourse*, volume 50. Springer Science & Business Media.

ASHER N., HUNTER J., MOREY M., FARAH B. & AFANTENOS S. (2016). Discourse structure and dialogue acts in multiparty dialogue : the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 2721–2727, Portorož, Slovenia : European Language Resources Association (ELRA).

ASHER N. & LASCARIDES A. (2003). *Logics of conversation*. Cambridge University Press.

ATWELL K., LI J. J. & ALIKHANI M. (2021). Where are we in discourse relation recognition ? In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, p. 314–325.

BADENE S., THOMPSON K., LORRÉ J.-P. & ASHER N. (2019). Data programming for learning discourse structure. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 640–645, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1061.

BARZILAY R. & LAPATA M. (2008). Modeling local coherence : An entity-based approach. *Computational Linguistics*, **34**(1), 1–34.

BHATIA P., JI Y. & EISENSTEIN J. (2015). Better document-level sentiment analysis from RST discourse parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 2212–2218, Lisbon, Portugal : Association for Computational Linguistics. DOI : 10.18653/v1/D15-1263.

CHEN J., LI X., ZHANG J., ZHOU C., CUI J., WANG B. & SU J. (2020). Modeling discourse structure for document-level neural machine translation. In *Proceedings of the First Workshop on Automatic Simultaneous Translation*, p. 30–36.

CHEN J. & YANG D. (2021). Structure-aware abstractive conversation summarization via discourse and action graphs. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1380–1391, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.109.

CHEN M., CHU Z. & GIMPEL K. (2019). Evaluation benchmarks and learning criteria for discourse-aware sentence representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 649–662.

CHI T.-C. & RUDNICKY A. (2022). Structured dialogue discourse parsing. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, p. 325–335.

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

EISNER J. (1996). Three new probabilistic models for dependency parsing : An exploration. In *COLING 1996 Volume 1 : The 16th International Conference on Computational Linguistics*.

GESSLER L., BEHZAD S., LIU Y. J., PENG S., ZHU Y. & ZELDES A. (2021). Discodisco at the disrpt2021 shared task : A system for discourse segmentation, classification, and connective detection. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, p. 51–62.

HUBER P. & CARENINI G. (2022). Towards understanding large-scale discourse structures in pre-trained and fine-tuned language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*.

KARIMI H. & TANG J. (2019). Learning hierarchical discourse-level structure for fake news detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 3432–3442.

LASCARIDES A. & ASHER N. (1993). Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and philosophy*, **16**(5), 437–493.

LEWIS M., LIU Y., GOYAL N., GHAZVININEJAD M., MOHAMED A., LEVY O., STOYANOV V. & ZETTLEMOYER L. (2020). BART : Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7871–7880, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.703.

LI C. (2023). *Facing Data Scarcity in Dialogues for Discourse Structure Discovery and Prediction*. Thèse de doctorat, Université de Lorraine.

LI C., HUBER P., XIAO W., AMBLARD M., BRAUD C. & CARENINI G. (2023). Discourse structure extraction from pre-trained and fine-tuned language models in dialogues. In *Findings of the Association for Computational Linguistics : EACL 2023*, p. 2517–2534.

LI J., LIU M., KAN M.-Y., ZHENG Z., WANG Z., LEI W., LIU T. & QIN B. (2020). Molweni : A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 2642–2652, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.238.

LIU Y. & LAPATA M. (2018). Learning structured text representations. *Transactions of the Association for Computational Linguistics*, **6**, 63–75. DOI : 10.1162/tacl_a_00005.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*.

LIU Z. & CHEN N. (2021). Improving multi-party dialogue discourse parsing via domain integration. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, p. 122–127, Punta Cana, Dominican Republic and Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.codi-main.11.

LOWE R., POW N., SERBAN I. & PINEAU J. (2015). The Ubuntu dialogue corpus : A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, p. 285–294, Prague, Czech Republic : Association for Computational Linguistics. DOI : 10.18653/v1/W15-4640.

QIN K., WANG L. & KIM J. (2017). Joint modeling of content and discourse relations in dialogues. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 974–984, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/P17-1090.

SHI W. & DEMBERG V. (2019). Next sentence prediction helps implicit discourse relation classification within and across domains. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, p. 5790–5796.

SHI Z. & HUANG M. (2019). A deep sequential model for discourse parsing on multi-party dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, p. 7007–7014.

WANG A., SONG L., JIANG H., LAI S., YAO J., ZHANG M. & SU J. (2021). A structure self-aware model for discourse parsing on multi-party dialogues. In *Proceedings of the Thirtieth International Conference on International Joint Conferences on Artificial Intelligence*.

Actes des journées LIFT 2023
Nancy, France, 20-21 novembre 2023

Can LLMs be used to understand clinical notes better?

Aman Sinha[1,2,4] Cristina Garcia Holgado[3]
Marianne Clausel[1] Mathieu Constant[2] Xavier Coubez[4]
(1) IECL, Université de Lorraine, Nancy, France
(2) ATILF, Université de Lorraine, Nancy, France
(3) Université de Poitiers, France (4) ICANS, Strasbourg, France
aman.sinha@univ-lorraine.fr

ABSTRACT

Clinical narrative plays important role to understand the patient's history and is instrumental for the clinicians to take decisions. Clinical text as general language reflects diverse author demographics and backgrounds. The variation found in the clinical narrative can lead the models to pick up on undesirable learning signals. Recently, large language models (LLMs) have demonstrated impressive capabilities at clinical knowledge. Although, they remain inferior to clinicians, LLMs show exciting potential applications in medicine, including knowledge retrieval, clinical decision support, etc. In this work, we are investigating the use of LLMs to overcome the impact of writing variation in clinical notes for contextual classification of clinical events.

RÉSUMÉ

Les LLMs peuvent-ils être utilisés pour mieux comprendre les notes cliniques ?

Le récit clinique joue un rôle important dans la compréhension de l'histoire du patient et aide les cliniciens à prendre des décisions. Le texte clinique, en tant que langage général, reflète la diversité de la démographie et des antécédents des auteurs. La variation observée dans le récit clinique peut conduire les modèles à capter des signaux d'apprentissage indésirables. Récemment, les LLMs ont démontré des capacités impressionnantes, notamment en matière de connaissances cliniques. Bien qu'ils restent inférieurs aux cliniciens, les LLMs présentent des applications potentielles intéressantes en médecine, notamment pour la recherche de connaissances, l'aide à la décision clinique, etc. Dans ce travail, nous étudions l'utilisation des LLMs pour surmonter l'impact de la variation de l'écriture dans les notes cliniques pour la classification contextuelle des événements cliniques.

MOTS-CLÉS : Notes cliniques ; LLMs ; Traitement du langage naturel.

KEYWORDS: Clinical Notes ; LLMs ; Natural Language Processing.

1 Introduction

Clinical notes help clinicians in answering prognostic inquiries and are essential for indentifying the confounding interactions between disease progression and interventions (Pham

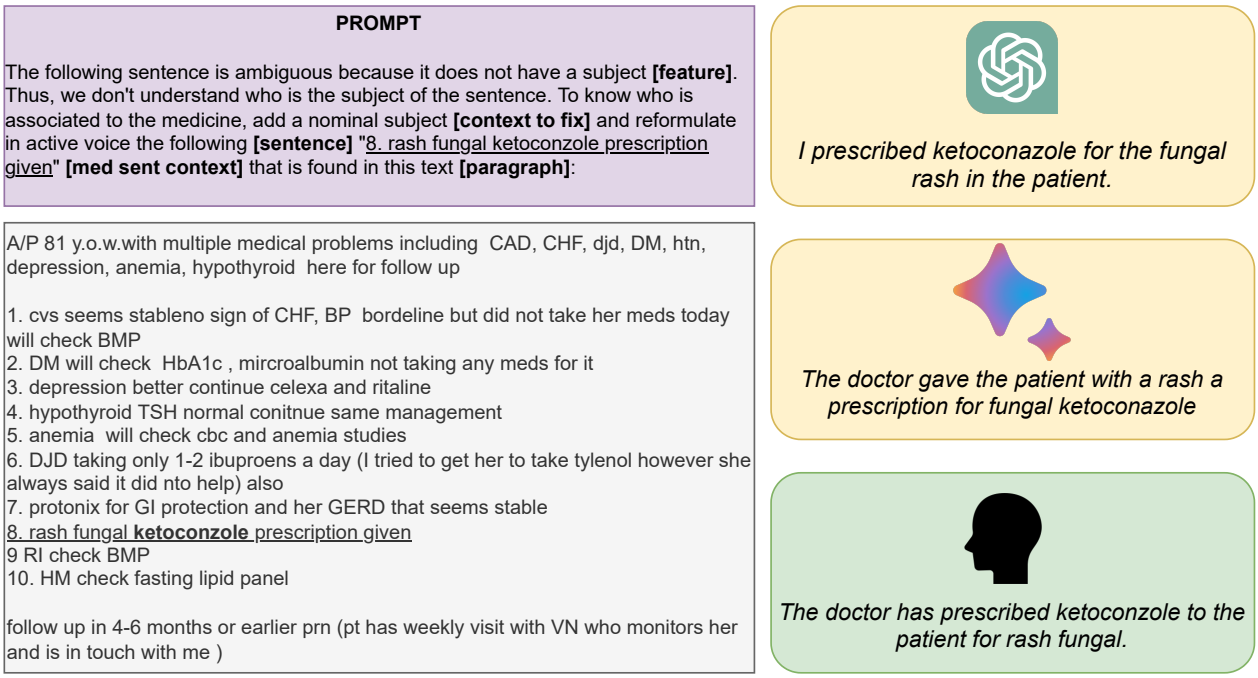


FIGURE 1 – Visual illustration for explanation generation via LLMs. Bottom-left box is an example of a text-prompt for a **target medicine** in the underlined sentence in a clinical note.

et al., 2017). As the narrative found in clinical notes is influenced by the writing style of the clinical staff, the variation found in clinical narrative can lead models to pick up on patterns that do not generalize to other author-demographics, or to rely on undesirable word-demographics relationships (Shah *et al.*, 2020). Such variations can be characterized by various cues such as the use of subject, shorthand descriptors, formulation of complex sentences, etc. Recently, LLMs have shown exciting assisting applications in various fields and they show great promise for applications in medicine (Singhal *et al.*, 2023) such as clinical decision support systems. In this work, we investigate the use of LLMs as an explanation agent to help the model to better understand clinical notes for context classification of *medication change* event.

2 Methodology

2.1 Experimental Setup

Dataset The CMED (Mahajan *et al.*, 2022) dataset was constructed using the 2014 i2b2/ UTHealth NLP shared task corpus. The clinical notes for this corpus consists of patients who were admitted for the Type II Diabetes (T2D). It contains 500 clinical notes which were annotated for medication extraction and for context annotations.

Contextual Classification Task We consider the context classification task from the n2c2 challenge 2022 [1] for identifying the contextual information associated to any *medication*

1. https ://n2c2.dbmi.hms.harvard.edu/2022-track-1

Sentences (CMED)	*linguistic feature*	LLM generated explanation
...she is amenable to a trial of tapering off the **Dilantin**. **Pravachol** 40 mg.	*has_subj* : Yes *has_subj* : No	- The doctor prescribed **Pravachol** 40 mg. at bedtime. She is tolerating the Pravachol well.
nitroglycerin yesterday, today relieved after 20 minutes with 1 .. Started on **oxygen**, lasix, combivent nebs.	*context_meds* : 0 *context_meds* : 2	- Oxygen : Patient was started on **oxygen** therapy. Lasix (furosemide) : Patient was started on Lasix, a diuretic, to help remove excess fluid from the body. Combivent nebs : Patient was started on Combivent nebulizers, a bronchodilator, to help open up the airways and improve breathing.
A UA was found to be positive and the pt was started on **Levo**. ..stay and resumption of **warfarin** if there are no contraindications	*complex_sent* : No *complex_sent* : Yes	- If there are no contraindications, stay and resume **warfarin**.

TABLE 1 – Sample linguistic cues in sentences containing target medicine (in **bold**)

change mentioned in the clinical note. For example, in the sentence ".. *He will not take his Glyburide the night before the surgery and the morning of* ..", for the underlined target medicine (i.e. Glyburide) we are interested to answer the following questions : Action (*What is the change being discussed ?* : [Start, Stop, Increase, Decrease, OtherChange, UniqueDose, Unknown]) ; Negation (*Is the change being discussed negated ?* : [Negated, NotNegated]) ; Temporality (*When is this change intended to occur ?* : [Past, Present, Future, Unknown]) ; Certainity (*How likely is this change to have occured/ will occur ?* : [Certain, Hypothetical, Conditional, Unknown]) ; and, Actor (*Who initiated the change ?* : [Physician, Patient, Unknown])

Contextual Classification Model We consider `cme2net` model (Sinha *et al.*, 2022) to analyze the sensitivity of the model for context classification task when trained on different linguistic cues explanation datasets. The `cme2net` model comprises of two encoders namely, static embeddings and PLM-based embeddings modules, a feature concatenation layer and two linear HEAD layers. The model is trained jointly with an additive weighted cross-entropy for NER+Event classification and context classification. As we are only interested in the context classification task, we use disable the first HEAD, use the annotations for NER+Event classification and only retrain the second HEAD for our study.

LLM as a helper We generate explanations for target medicine with respect to any linguistic cue by providing a pre-designed text-prompt along with the sentence and the context associated to the target medicine. In Figure 1, we show a sample of the text-prompt (in purple box, top left) for LLM based chat assistants (such as ChatGPT/Google Bard). The actual prompt is given in the gray box (left bottom) in which the sentence containing the target medicine is underlined and the target medicine is put in bold. On the right, we show the explanation responses obtained from the two LLM agents and a manual response.

In Table 1, we provide examples of generated explanations for different linguistic cues such as `has_subj` (if the subject associated to the target medicine is mentioned in the sentence), `context_meds` (if there are other medicines mentioned around the target medicine) and

`complex_sent` (if the sentence containing the target medicine is complex in nature). We use publicly accessible GPT-3.5 Turbo (OpenAI, 2023) API as the explanation agent in our study and create explanation datasets for each linguistic cues separately.

2.2 Our approach

For any sentence s with a target medicine m, and the context c we examine a linguistic feature $l \in \text{L} = \{\varnothing, l_1, l_2, ..\}$ in a dataset denoted by D. We refer to `cme2net` (Sinha *et al.*, 2022) which is the context classification model as CC and LLMs (OpenAI, 2023) (such as ChatGPT) as *llm*. We perform the steps below for our study :

1. Firstly, train `cme2net` model on original dataset ($D(\varnothing) \equiv D$). Collect $P_{D(\varnothing)} \equiv P_{\varnothing}$ which refers to test predictions for no *modified* linguistic feature.
2. For each $l \in \text{L} \setminus \{\varnothing\}$:
   (a) Divide dataset into two parts : $D(l)$ and $D(\neg l)$ (*read D w/o linguistic feature l*)
   (b) For $D(\neg l)$, generate *llm* explanations, and create D^{llm} such that it contains l.
   (c) Obtain `cme2net`* model trained on $D(l)$+D^{llm}. Collect test predictions P_l
   (d) Compare Q($P_{\varnothing}$) and Q(P_l), where Q(.) is the distribution.
3. Calculate *overall* statistical significance to verify if L has an impact on CC.

3 Discussion & Future Work

Preliminary Evidence The overview (Mahajan *et al.*, 2023) summarizes various scenarios where the different participant models fail on context classification task. These cases include (i) when target medicine is associated with more than one set of contexts, (ii) use of shorthand descriptors such as AC (i.e., before meal), Qd (i.e., daily), medications with short names (such as Ca, K, O2, EPO, etc.), (iii) when target medicine is associated to Conditional/Hypothetical where longer textual cues are needed, and (iv) no action verb is mentioned associated to target medicine.

Our analysis We conducted our analysis on the predictions obtained from `cme2net` model and found that out of 328, model struggled in 195 cases where there was no clear person mentioned associated to the target medicine. Further, out of a total 328 cases, the model struggled in 273 cases where the sentence was complex (ambiguous).

Future Work Our analysis clearly in line with the observation made by (Mahajan *et al.*, 2023) about model's behaviour sensitivity when subjected to certain linguistic cues (for eg. presence of subject, listing of medicines). Further, we would like to apply our LLM based approach to examine model's sensitivity towards different linguistic cues and explore model's sensitivity towards them.

Références

MAHAJAN D., LIANG J. J. & TSOU C.-H. (2022). Toward Understanding Clinical Context of Medication Change Events in Clinical Narratives. *AMIA Annual Symposium Proceedings*, **2021**, 833–842.

MAHAJAN D., LIANG J. J., TSOU C.-H. & UZUNER Ö. (2023). Overview of the 2022 n2c2 shared task on contextualized medication event extraction in clinical notes. *Journal of Biomedical Informatics*, p. 104432.

OPENAI (2023). Gpt-4 technical report. *ArXiv*, **abs/2303.08774**.

PHAM T., TRAN T., PHUNG D. & VENKATESH S. (2017). Predicting healthcare trajectories from medical records : A deep learning approach. *Journal of biomedical informatics*, **69**, 218–229.

SHAH D. S., SCHWARTZ H. A. & HOVY D. (2020). Predictive biases in natural language processing models : A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 5248–5264, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.468.

SINGHAL K., AZIZI S., TU T., MAHDAVI S. S., WEI J., CHUNG H. W., SCALES N., TANWANI A., COLE-LEWIS H., PFOHL S. *et al.* (2023). Large language models encode clinical knowledge. *Nature*, **620**(7972), 172–180.

SINHA A., VISHWAKARMA A., CLAUSEL M. & CONSTANT M. (2022). Cme^2 net : Contextual medical event extraction network for clinical notes. *Proceedings of the 19th International Conference on Natural Language Processing (ICON) : Workshop on Context-aware NLP in eHealth*, p. 1–7.

Actes des journées LIFT 2023
Nancy, France, 20-21 novembre 2023

Création semi-automatique d'un lexique bilingue langue des signes française (LSF) / français pour l'annotation de vidéos de LSF

Julie Lascar Annelies Braffort Michèle Gouiffès
Université Paris-Saclay, CNRS, LISN
Campus Universitaire Bat 507, rue du Belvédère, 91405 Orsay, France
prenom.nom@lisn.upsaclay.fr

RÉSUMÉ

Cet article présente nos contributions sur la constitution de ressources et le traitement automatique au service de l'analyse de vidéos de langue des signes française (LSF) et plus particulièrement de l'annotation automatique des unités lexicales. À ce jour, nous avons pu constituer de manière semi-automatique un lexique bilingue comportant 88 entrées et utilisé pour élaborer un classificateur qui a permis d'annoter automatiquement des unités lexicales sur le corpus Mediapi-rgb.

ABSTRACT

Semi-automatic creation of a bilingual French sign language (LSF) / French lexicon for annotating LSF videos

This article presents our contributions to the constitution of resources and automatic processing for the analysis of French sign language (LSF) videos, and more specifically the automatic annotation of lexical units. To date, we have been able to build a bilingual lexicon containing 88 entries in a weakly supervised manner. This lexicon has been used to build a classifier, used to automaticaly annotated lexical units on the Mediapi-rgb corpus.

MOTS-CLÉS : Langue des signes française, lexique bilingue, annotation automatique.

KEYWORDS: French sign language, bilingual lexicon, automatic annotation.

1 Introduction

Les Langues des Signes (LS) sont des langues naturelles pratiquées au sein des communautés de Sourds et la Langue des Signes Française (LSF) est celle pratiquée en France. À ce jour, les LS sont encore très peu dotées et les recherches sur ces langues sont assez récentes, en particulier dans le domaine du traitement automatique. Nos projets actuels ont pour objectif de contribuer à la constitution de ressources et de traitements automatiques au service de l'analyse de vidéos de LSF. Cet article présente un point d'étape sur nos contributions autour de l'annotation automatique.

FIGURE 1 – Capture d'écran du site Média'Pi !

2 Constitution d'un lexique bilingue

Il n'existe pas à l'heure actuelle de lexique bilingue en contexte exploitable pour le traitement automatique de la LSF. D'une manière générale, les ressources de LSF utilisables pour le traitement automatique sont peu nombreuses (Braffort, 2022). Une précédente étude (Bull, 2023) a permis de constituer un jeu de données utilisable pour le traitement automatique. Il a été constitué à partir du corpus Mediapi-rgb, comportant 86h de vidéos en LSF produites par des journalistes ou présentateurs sourds du média bilingue en ligne Média'Pi ! [1], accompagnées de sous-titrage en français (figure 1).

À partir de ce jeu de données, nous avons constitué un premier lexique bilingue avec une approche faiblement supervisée :

- nous avons sélectionné une liste de mots présents dans les sous-titres de telle sorte que : 1) à part quelques exceptions, ils possèdent peu, ou pas de synonymes et 2) leur équivalent en LSF varie peu en fonction du contexte. Nous avons ainsi opté pour les jours, les mois, certaines villes et pays ainsi que du vocabulaire lié à l'actualité de l'époque (masque, chômage, gilet jaune...) ;
- pour chaque mot, nous avons sélectionné toutes les vidéos pour lesquelles le sous-titre contient ce mot, puis nous avons calculé la similarité entre paires de vidéos (méthode de la similarité cosinus) ;
- dans chaque vidéo, on retient les numéros d'images pour lesquelles la similarité dépasse un certain seuil (empiriquement fixé à 0.6) sur un nombre consécutif d'au moins 4 images.

1. https ://www.media-pi.fr/

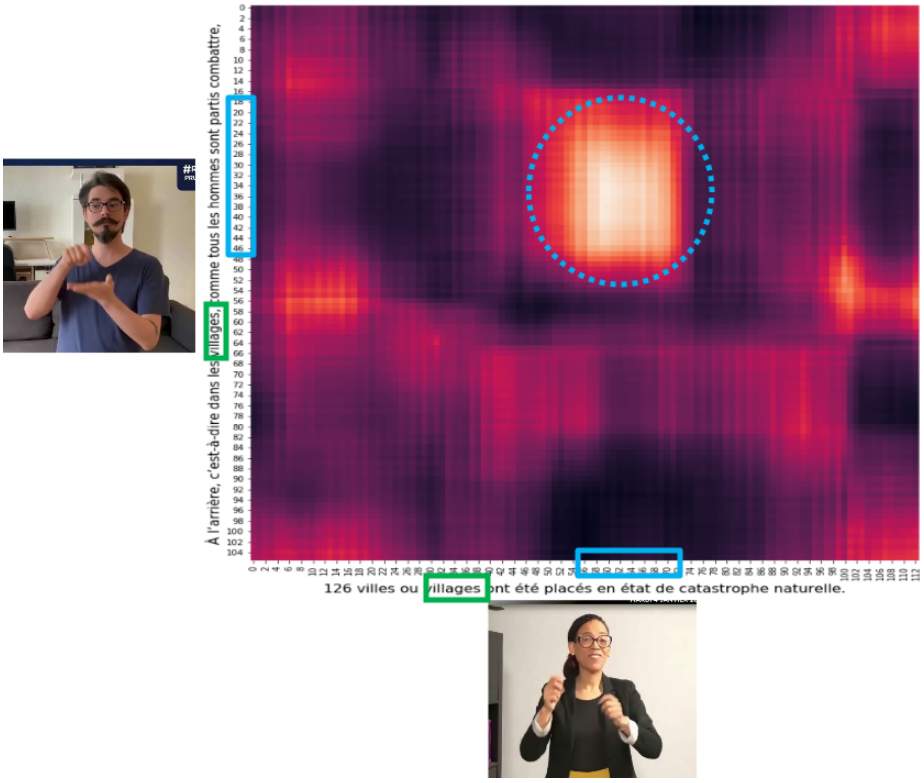


FIGURE 2 – Exemple de matrice de similarité entre deux vidéos pour le mot « village ».

La figure 2 illustre une matrice de similarité entre deux vidéos dont les sous-titres comportent le mot « village ». La zone plus claire est celle pour laquelle la similarité est la plus forte.

Certains signes peuvent présenter des variantes de forme qui dépendent du locuteur. Par exemple quelques signes représentant les mois sont réalisés par des locuteurs avec une main et par d'autres avec les deux mains. Pour distinguer ces variantes, nous avons procédé à un regroupement automatique des vidéos par locuteur. Par ailleurs, certains mots peuvent présenter des variantes de sens, comme par exemple le mot « place » qui est signé différemment selon son sens (ex : mettre en place, quatrième place, place de la Concorde). Nous avons utilisé un modèle de language Bert (Devlin *et al.*, 2019) pour regrouper les mots selon leur sens dans le contexte de la phrase.

Une dernière étape a consisté à regrouper les vidéos capturées pour chaque mot en utilisant une méthode de partitionnement des données (algorithme des k-moyennes). Cela a permis d'une part, d'éliminer les erreurs de détection et d'autre part, de détecter certaines variantes, comme illustré sur la figure 3.

Dans ces figures, après réduction de la dimension des données par analyse par composantes principales (ACP), on a projeté sur les deux axes principaux les séquences obtenues pour les signes ITALIE et NOVEMBRE. Dans la figure de gauche, on obtient deux groupes : le plus grand regroupe les vidéos correspondant effectivement au signe ITALIE, le plus petit regroupe les erreurs de détection. Dans la figure de droite, on obtient aussi deux groupes

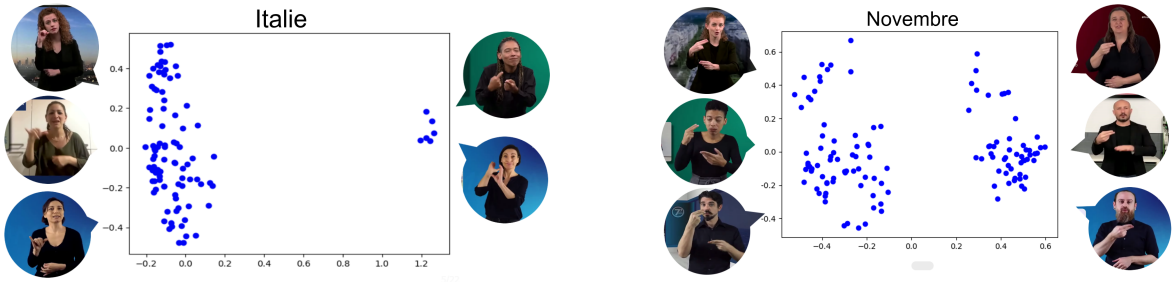


FIGURE 3 – Groupes obtenus pour les signes ITALIE et NOVEMBRE (après réduction de la dimension des données par ACP).

qui permettent de différencier deux variantes pour le signe NOVEMBRE : celui de gauche regroupe les signes effectués à deux mains et celui de droite regroupe les signes effectués à une seule main.

Nous obtenons ainsi des classes associant forme et sens distinctes et présentant peu d'erreurs. Une dernière phase de contrôle réalisée par des experts est en cours et va permettre de ne garder que des occurences correctes et bien segmentées, c'est-à-dire pour lesquelles la séquence d'images comporte l'ensemble du signe et aucun élément de coarticulation avec les signes précédent et suivant.

3 Constitution d'un classifieur basé sur ce lexique

La constitution de ce lexique bilingue a permis d'annoter une partie des données d'entraînement du corpus Mediapi-rgb. Ces données annotées ont ensuite été utilisées pour entraîner un classifieur, qui prend en entrée les vidéos sous forme de séquences d'images et produit en sortie des séquences d'entiers, chaque entier identifiant pour chaque image la classe correspondante. L'architecture du système, illustrée figure 4, comporte deux modèles :

- Le premier a pour rôle d'encoder les vidéos sous forme de plongements.
- Le second est le classifieur qui prend en entrée ces plongements .

Des séries d'expériences ont été menées en faisant varier les modèles utilisés. Pour l'étape de calcul des plongements, trois modèles simples ont été comparés, tous adaptés au fait que les données sont de nature temporelle : un modèle de type *transformers* entraîné sur des données d'actions humaines (*Video Swin Transformer* (Liu *et al.*, 2022)), un modèle à base de réseaux de convolution 3D (I3D) réentraîné sur un corpus de BSL (langue des signes britanique) (Renz *et al.*, 2021) et de nouveau le modèle *Video Swin Transformer* mais cette fois-ci réentrainé sur des vidéos de BSL (Prajwal *et al.*, 2022). Le troisième modèle s'avère le plus performant et c'est celui que nous avons conservé pour la suite.

Pour l'étape de classification plusieurs modèles ont été comparés (des perceptrons et des LSTM, à une ou deux couches). Nous avons optimisé les paramètres en utilisant le score F1, qui est la moyenne harmonique entre la précision et le rappel. Plus précisément, chaque

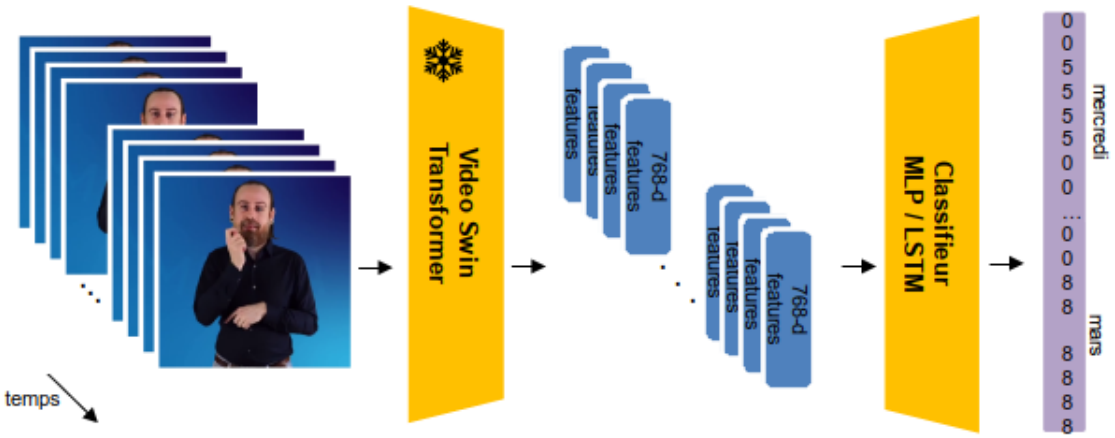


FIGURE 4 – Architecture du classifieur.

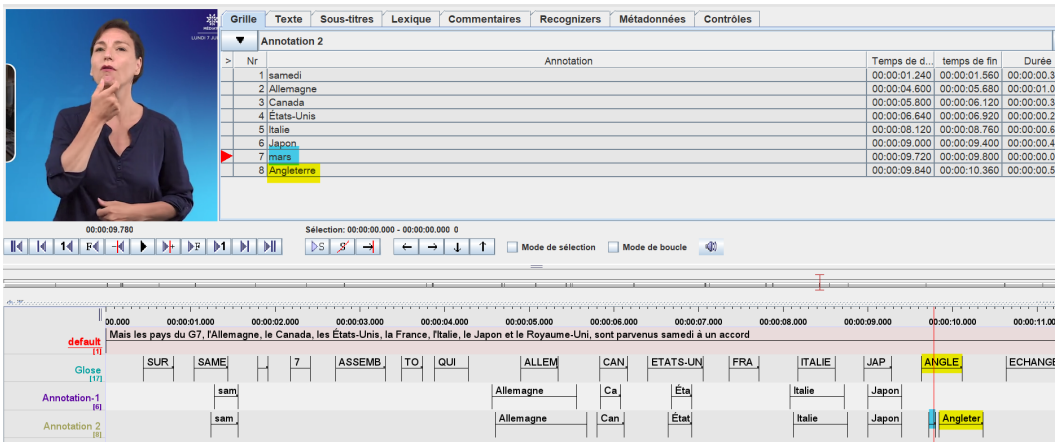


FIGURE 5 – Annotation par des experts avec Elan.

classe se voit attribuer un score F1 (voir l'annexe pour le détail du calcul). Le score final correspond à la moyenne des scores F1 de chaque classe. Nous avons retenu une architecture LSTM à une couche qui donne le meilleur score (score F1 : 0.79).

Ce classifieur a ensuite été utilisé pour annoter de nouveau les données d'entraînement. Ainsi, 5 876 vidéos et 7 259 signes ont été annotés automatiquement. Le score F1 est insuffisant pour évaluer la qualité des annotations, en particulier sur la qualité de la segmentation temporelle. Pour cette étape aussi, une analyse fine des résultats par des experts est en cours. La figure 5 montre un exemple d'annotation à l'aide du logiciel Elan (Crasborn & Sloetjes, 2008) des unités lexicales sur une vidéo en LSF accompagnée du sous-titrage « Mais les pays du G7, l'Allemagne, le Canada, les Etats-Unis, la France, l'italie, le Japon et le Royaume-Uni, sont parvenus samedi à un accord ».

L'annotation comporte trois pistes : *Glose* est renseignée manuellement par un expert, tandis que *Annotation-1* et *Annotation-2* sont renseignées automatiquement, respectivement suite à l'étape 1 basée sur le calcul de similarité et à l'étape 2 à l'aide du classifieur.

Sur cet exemple, l'ensemble des signes présents dans le lexique ont été détectés à l'exception du signe FRANCE et du signe ANGLETERRE dans l'étape 1. L'étape 2 améliore les résultats car le signe ANGLETERRE a été détecté. Cependant un segment glosé MARS a été inséré. Il se trouve que les signes MARS et ANGLETERRE ont des formes proches (configuration et emplacement de la main). Enfin, globalement, les annotations automatiques ont bien positionné les segments dans le flux temporel mais ne sont pas très précis. L'expertise en cours sur la qualité des occurrences du lexique bilingue (construites lors de l'étape 1) montre qu'un grand nombre d'occurrences sont incomplètes ou incluent des parties des coarticulations avec les signes précédent et suivant. L'amélioration de la qualité du lexique bilingue devrait significativement améliorer les résultats du classifieur, y compris sur l'aspect repérage temporel.

4 Bilan et perspectives

Il s'agit des tous premiers pas vers un système d'annotation automatique d'unités lexicales dans des vidéos de LSF et beaucoup reste à faire sur l'exploitation des sous-titres, la représentation des vidéos, l'architecture du classifieur et l'évaluation des performances du système. Ces différents axes sont actuellement en chantier, ainsi que l'augmentation du vocabulaire, qui dépasse maintenant les 1000 classes. Nous envisageons aussi d'exploiter les annotations fines réalisées par des experts sur le corpus Dicta-Sign-LSF-V2 (Belissen *et al.*, 2020), afin de récupérer des occurrences de bonne qualité et ajouter de nouvelles classes. Nous envisageons aussi d'évaluer le classifieur sur ce corpus.

Par la suite, ce travail doit être étendu à d'autres unités gestuelles telles que les structures illustratives très fréquentes dans les LS et qui ne peuvent pas être listées dans un dictionnaire, ce qui implique de concevoir une approche différente.

Références

BELISSEN V., BRAFFORT A. & GOUIFFÈS M. (2020). Dicta-Sign-LSF-v2 : Remake of a continuous French Sign Language dialogue corpus and a first baseline for automatic sign language processing. In *Language Resources and Evaluation Conference*, p. 6040–6048, Marseille, FR : ELRA.

BRAFFORT A. (2022). Langue des Signes Française : Etat des lieux des ressources linguistiques et des traitements automatiques. In *Journées Jointes des GRD LIFT et TAL*, p. 131–138, Marseille, FR : CNRS.

BULL H. (2023). *Learning sign language from subtitles*. Thèse de doctorat. Université Paris-Saclay.

CRASBORN O. & SLOETJES H. (2008). Enhanced elan functionality for sign language corpora. In *Work. on the Representation and Processing of Sign Languages @ LREC Conf.*, p. 39–43, Marrakech, Morocco.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). Bert : Pre-training of deep bidirectional transformers for language understanding. In *Conf. of the NAACL association : Human Language Technologies*, p. 4171–4186 : ACL.

LIU Z., NING J., CAO Y., WEI Y., ZHANG Z., LIN S. & HU H. (2022). Video swin transformer. In *Conf. on Computer Vision and Pattern Recognition*, p. 3202–3211, New Orleans, USA : IEEE.

PRAJWAL K. R., BULL H., MOMENI L., ALBANIE S., VAROL G. & ZISSERMAN A. (2022). Weakly-supervised fingerspelling recognition in british sign language videos. In *British Machine Vision Conference*, London, UK : BMVA.

RENZ K., STACHE N. C., ALBANIE S. & VAROL G. (2021). Sign language segmentation with temporal convolutional networks. In *Conf. on Acoustics, Speech and Signal Processing*, p. 2135–2139, Toronto, Canada : IEEE.

Annexe : calcul du score F1 d'une classe

On obtient le score F1 d'une classe c de la manière suivante :

a. On calcule le score F1 pour chaque vidéo dans laquelle la classe c est présente en comparant les prédictions du modèle (Pred) aux annotations effectuées lors de l'étape 1 (GT).

$$\text{Vidéo 1} \rightarrow \begin{cases} \text{GT : 000111111000...} \\ \text{Pred : 011111000000...} \end{cases} \qquad \text{F1} = 0.55$$

$$\text{Vidéo 2} \rightarrow \begin{cases} \text{GT : 000111111000...} \\ \text{Pred : 000001110000...} \end{cases} \qquad \text{F1} = 0.75$$

...

b. On calcule la moyenne des scores F1 de chaque vidéo.

Actes des journées LIFT 2023
Nancy, France, 20-21 novembre 2023

Descripteurs Linguistiques et Caractérisation Objective des Catégories Textuelles

Marina Seghier[1] Alice Millour[1] Jean-Yves Antoine[2]
(1) LIASD - Université Paris 8, 2 rue de l'Université, 93526 Saint-Denis, France
(2) LIFAT - Université de Tours, 64 avenue Jean Portalis, 37200 Tours, France
ms@up8.edu, am@up8.edu, jean-yves.antoine@univ-tours.fr

RÉSUMÉ

Nous présentons de premiers résultats s'inspirant des travaux de Biber (1988) utilisant des descripteurs linguistiques pour redéfinir les dimensions de la variation textuelle en français.

ABSTRACT

Linguistic Descriptors and Objective Characterization of Textual Categories

We present first results inspired by the work of Biber (1988) using linguistic descriptors to redefine the dimensions of textual variation in French.

MOTS-CLÉS : variation textuelle, classification non supervisée, descripteurs linguistiques, évaluation, annotation.

KEYWORDS: textual variation, unsupervised classification, linguistic descriptors, evaluation, annotation.

Les systèmes développés et de plus en plus répandus aujourd'hui, sont présentés comme étant très performant pour un grand nombre de tâches. Or, les performances annoncées ne sont pas toujours celles rencontrées selon les types (ou "genres") de ressources textuelles auxquels ces outils sont confrontés. En effet, on peut par exemple observer un différentiel de performances important entre différentes catégories de texte pour la tâche de reconnaissance d'entités nommées (Millour *et al.*, 2022) et d'étiquetage morpho-syntaxique. Cependant, les typologies textuelles existantes, fondées sur une classification des catégories *a priori* sans justification linguistique, ne permettent pas d'expliquer ce différentiel.

1 Précédents dans la Classification Textuelle

Biber (1988) a été le premier à mener une étude statistique sur ce qu'il appelle "genre" en anglais britannique, dans le but d'identifier plusieurs dimensions de la variation dans la langue. Il a procédé à une "analyse factorielle" (ACP) sur le LOB (*Lancaster-Oslo-Bergen*, un million de mots environ – composé de plusieurs échantillons de "genres" issus de l'écrit :

presse, religion, humour...), le *London-Lund* (500 000 mots environ – composé de textes de parole transcrite : entretiens, discours spontanés, préparés...), et une collection de ses propres lettres manuscrites.

Au terme de son analyse basée sur 67 caractéristiques calculées à partir de descripteurs linguistiques (tels que les pronoms personnels, les verbes au passé, la longueur des mots, des phrases, etc.), il a affirmé que la variation linguistique était continue selon six dimensions : 1) impliqué (affectivement) *VS* informationnel, 2) narratif *VS* non-narratif, 3) référence explicite *VS* dépendante de la situation, 4) expression manifeste de persuasion, 5) abstrait *VS* non-abstrait et 6) production d'informations sous contrainte temporelle.

À la manière de Passonneau *et al.* (2014), qui ont actualisé cette recherche pour l'anglais américain à partir du corpus MASC (*Manually Annotated Sub-Corpus*, 500 000 mots environ), nous proposons une caractérisation linguistique de différents genres à partir d'un corpus multisource pour le français.

2 Des Données à l'Analyse en Composantes Principales

Nous avons travaillé à partir de FENEC (*FrEnch Named-entity Evaluation Corpus*, 11 000 tokens environ – cf. table 1) (Millour *et al.*, 2022), un corpus d'évaluation pour la tâche de reconnaissance d'entités nommées en français. Il est composé de onze documents de six catégories textuelles différentes (poésie, prose, parole transcrite, encyclopédie, informations, multi-sources) et annoté manuellement en entités nommées selon divers jeux d'étiquettes.

Document	Période	Genre	Nb. phrases (Nb. tokens)	Licence
42131-0 (*Traité sur la Tolérance*, Voltaire)	XVIIIe	prose	40 (1 020)	Project Gutenberg
pg6470 (*Le Ventre de Paris*, Émile Zola)	XIXe		51 (1 002)	Project Gutenberg
pg6099 (*Les Fleurs du Mal*, Baudelaire)	XIXe	poésie	30 (1 014)	Project Gutenberg
56708-0 (*Œuvres d'Arthur Rimbaud - Vers et proses*)	XIXe		52 (1 027)	Project Gutenberg
UD French GSD	XXIe	multisources	35 (1 021)	CC BY-SA 4.0
Sequoia (Candito & Seddah, 2012)	XXIe		44 (1 002)	Licence LGPL-LR
French Question Bank (Seddah & Candito, 2016)	XXIe		102 (1 006)	Licence LGPL-LR
APIL (office du tourisme Othe-Armance)	XXIe	informations	29 (1 002)	Licence LGPL-LR
Wikinews	XXIe		46 (1 024)	CC BY 2.5
WikiNER français	XXIe	encyclopédie	36 (1 003)	CC BY 4.0
Spoken (Rhapsodie (Lacheret, Anne *et al.*, 2014))	XXIe	parole	70 (1 028)	CC BY-SA 4.0

TABLE 1 – Contenu du corpus annoté FENEC (échantillons de 1 000 tokens environ dans six genres) (Millour *et al.*, 2022).

Étant donné la taille peu conséquente de notre jeu de données, nous avons choisi de former des échantillons de 200 tokens, à partir desquels nous avons mesuré l'importance de certains descripteurs linguistiques pour le typage des textes, tels que les entités nommées, mais également les parties du discours (verbes, noms, déterminants, adjectifs...) et les verbes au

passé.

2.1 Choix et calcul des caractéristiques

Les caractéristiques utilisées pour cette expérience sont :

— les entités nommées du corpus FENEC (selon le schéma fin Quaero [1]), à savoir : PER, LOC, ORG, MISC (respectivement : personnes, lieux, organismes, divers) ;

— les parties du discours, pour lesquels nous avons testé plusieurs outils : 1) le modèle POET (*A French Extended Part-of-Speech Tagger*) de FLAIR [2], 2) CAMEMBERT [3] et 3) SPACY [4] ;

— les verbes au passé, annotés avec le *morphologizer* de SPACY [5].

Concernant l'étiquetage en parties du discours, après une brève comparaison des sorties produites par les trois outils, nous avons choisi de conserver les annotations du modèle POET de FLAIR pour calculer nos caractéristiques. Nous avons également réalisé une évaluation manuelle de ses résultats (environ 100 étiquettes par document) et de SPACY (toutes les occurrences de verbes annotés au passé, soit le rappel).

genre	Flair (précision)	SpaCy (rappel)
parole	78,43	58,14
encyclopedie	94,64	86,67
informations	90,29	77,42
prose	88,56	77,12
poesie	90,14	37,29

TABLE 2 – Résultats de l'évaluation manuelle de FLAIR et SPACY.

Sur la base de 26 caractéristiques calculées, nous avons choisi de mener notre expérience sur 23 d'entre elles, en retirant :

— MISC, soit les entités nommées "diverses" (*miscellaneous* en anglais) car ces dernières étaient trop variées au sein de cette catégorie (par exemple "boucher", "matelots", "enchanteresse", "*New York Times*", "*Hubble*", "*Surface and Depth*", "bataille d'Actium", "accords sur le charbon et l'acier"), et nous n'aurions pas été en mesure d'interpréter précisément et de manière fiable, l'impact de cette caractéristique sur l'ACP ;

1. Guide d'annotation : http://www.quaero.org/media/files/bibliographie/quaero-guide-annotation-2011.pdf
2. https://huggingface.co/qanastek/pos-french-camembert-flair
3. https://huggingface.co/gilf/french-camembert-postag-model
4. https://spacy.io/models/fr#fr_core_news_sm
5. https://spacy.io/models/fr

— X, soit les mots inconnus, tels que des mots anglais, des consonnes euphoniques ("t" dans "appelle-t-on"), "j" dans "j∼ j∼ j∼" (répétition dans la parole transcrite), car nous craignions un effet fort d'échantillonage, du fait du nombre total d'observables de cette étiquette dans le corpus, par rapport à la taille de celui-ci ;
— SYM, soit les symboles tels que le tilde, le pourcent, les symboles monétaires, l'abréviation de "environ" (env), pour la même raison que précédemment.

Dans la table 3, nous pouvons voir les occurrences d'un échantillon des observables pour 1 000 tokens.

	prose	parole	informations	encyclopedie	poesie
LOC	4	14	35	41	6
ORG	1	1	13	6	0
PER	23	5	8	22	11
TOTAL_EN	39	28	74	86	25
ADP	112	114	160	151	120
SCONJ	22	18	3	3	12
CCONJ	21	42	25	26	40
ADV	67	118	31	30	49
PROPN	23	27	66	93	17
NUM	10	19	54	57	3
AUX	37	42	22	35	19
VERB	125	135	75	66	82
DET	125	104	142	139	162
ADJ	110	118	123	137	156
NOUN	155	154	226	193	219
PRON	93	100	24	19	52
PPER1S	5	20	0	0	8
PPER2S	1	5	0	0	4
PPER3	25	16	8	7	7
INTJ	1	7	0	0	1
PUNCT	104	90	94	90	123
PAST_TENSE	59	43	36	45	30

TABLE 3 – Nombre d'occurrences d'un échantillon d'étiquettes pour 1 000 tokens, par genre.

2.2 Dimensions en français

D'après cette première analyse, une dimension semble se dessiner. D'une part, les textes de la catégorie 'poésie' sont davantage caractérisés par des mots plus longs et des noms détaillés, et précisés par des adjectifs ; ceux des catégories 'prose' et 'parole', par des conjonctions de subordination, des verbes en général, des verbes au passé, et à la voix passive. La présence d'entités nommées semble quant à elle caractériser davantage les genres 'informations' et 'encyclopédie'. En effet, nous pouvons supposer que les textes d'informations et d'encyclopédies présentent plus d'éléments tels que des zones géographiques, lieux d'intérêt, dates historiques, numéros de téléphones, etc.

Notre première composante dans la figure 1 présente des similitudes avec les cinquième

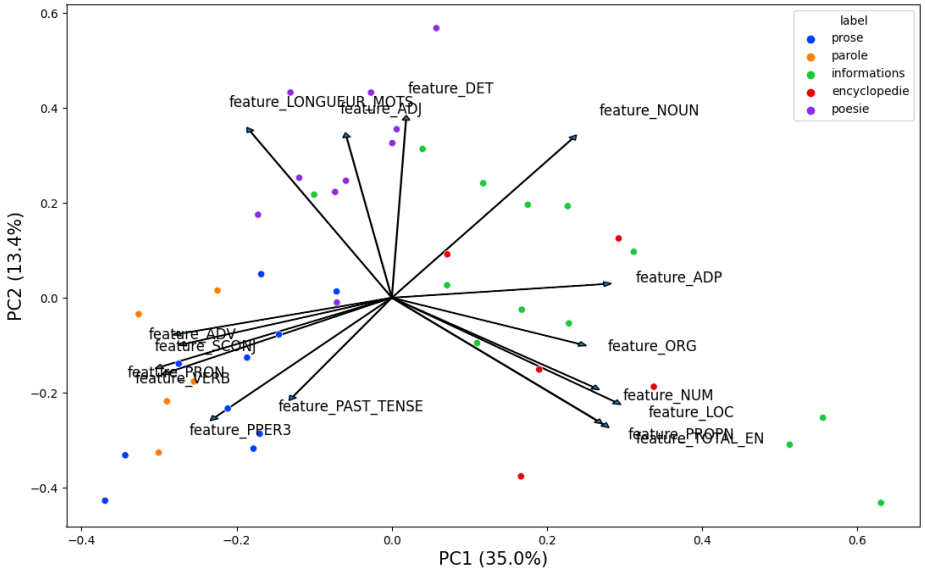


FIGURE 1 – Première et deuxième composantes principales, 23 caractéristiques.

de Biber (1988) et quatrième de Passonneau *et al.* (2014) (abstrait VS non-abstrait). Les catégories de productions écrites, détaillées d'un côté (poésie, prose) semblent plus conceptuelles (abstraites) ; et des catégories explicatives de l'autre (informations, encyclopédie) plus concrètes (non-abstraites).

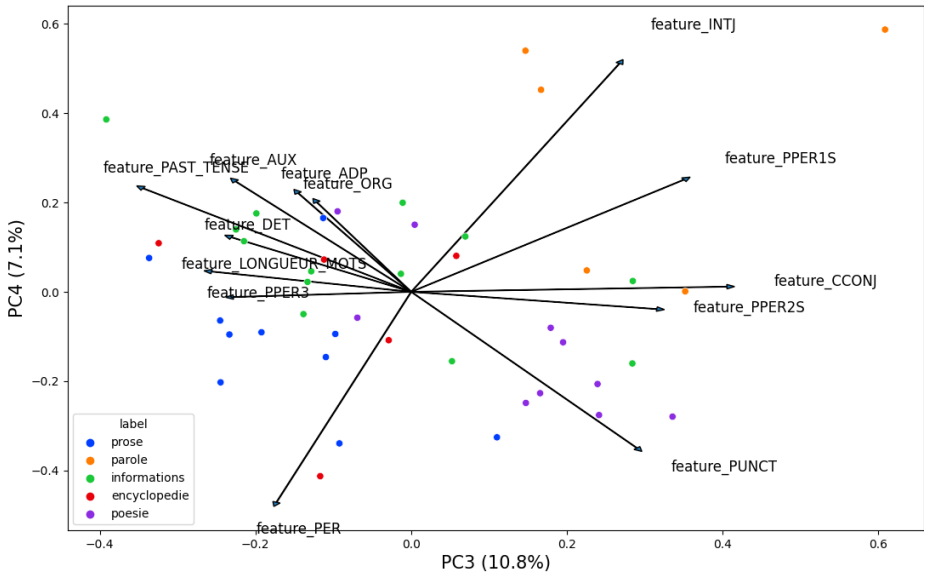


FIGURE 2 – Troisième et quatrième composantes principales, 23 caractéristiques.

D'après cette deuxième analyse, une dimension semble à nouveau se dessiner. D'une part, les textes de parole transcrite semblent caractérisés par des interactions à la 1ère et 2ème personne du singulier, et la présence d'interjections et de conjonctions de coordination. D'autre part, les textes en prose présentent davantage d'entités nommées de personnes, de pronoms à la 3ème personne et de ponctuation. Dans les textes d'informations, nous pouvons

trouver de la narration d'évènements passés, des mots plus longs et davantage de noms d'organismes.

Notre troisième composante dans la figure 2 présente ainsi des similitudes avec les deux premières de Biber (1988) et Passonneau *et al.* (2014) (impliqué VS non-impliqué). En effet, nous retrouvons d'un côté des textes d'interactions orales (parole) plus impliquées, et de l'autre des textes plus informationnels (encyclopédie, prose).

Contrairement aux deux premières composantes, celles-ci nous permettent de discriminer la parole de la prose, mais dans le cadre de cette étude, elles ne sont tout de même pas suffisantes pour discriminer plus nettement chaque catégorie textuelle, notamment celle de l'encyclopédie.

3 Conclusion

Bien qu'il reste une notion de continuum à approfondir, cette étude a permis de retrouver certaines similitudes avec les dimensions textuelles que Biber (1988) et Passonneau *et al.* (2014) avaient fait émerger en anglais britannique et américain, mais pour la première fois, sur un jeu de données multi-catégories français.

En outre, nous avons pu également éprouver trois outils : SpaCy pour à la fois l'annotation en parties du discours et en traits morphosyntaxiques, CamemBERT et le modèle POET de Flair pour l'annotation en parties du discours. L'évaluation de ces derniers nous a permis de constater sur ces deux tâches, ce qu'avaient observé Millour *et al.* (2022) sur la tâche de REN pour le français : un différentiel de performances important entre les catégories textuelles, avec des faiblesses majeures sur des textes de parole transcrite et de poésie.

Le cadre expérimental mis en place dans notre étude est modulable et permet facilement d'enrichir l'expérience avec d'autres caractéristiques, d'autres types de textes, et de donner lieu à de nombreuses extensions. L'une d'elles consiste à s'inspirer des travaux de Fu *et al.* (2020) et d'approfondir la notion des caractéristiques des entités nommées. Car au-delà de leurs types, ce sont sûrement leurs propriétés intrinsèques (longueur en caractères et en tokens, fréquence, ambiguïté, persistance...) qui mettent en difficulté les outils.

Références

BIBER D. (1988). *Variation across Speech and Writing*. Cambridge University Press. DOI : 10.1017/CBO9780511621024.

CANDITO M. & SEDDAH D. (2012). Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical (the sequoia corpus : Syntactic annotation and use for a parser lexical domain adaptation method) [in French]. In *Proceedings of*

the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN, p. 321–334, Grenoble, France : ATALA/AFCP.

FU J., LIU P. & NEUBIG G. (2020). Interpretable multi-dataset evaluation for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 6058–6069, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.489.

LACHERET, ANNE, KAHANE, SYLVAIN, BELIAO, JULIE, DISTER, ANNE, GERDES, KIM, GOLDMAN, JEAN-PHILIPPE, OBIN, NICOLAS, PIETRANDREA, PAOLA & TCHOBANOV, ATANAS (2014). Rhapsodie : un treebank annoté pour l´étude de l´interface syntaxe-prosodie en français parlé. *SHS Web of Conferences*, **8**, 2675–2689. DOI : 10.1051/shsconf/20140801305.

MILLOUR A., DUPONT Y., JOUGLAR A. & FORT K. (2022). FENEC : un corpus équilibré pour l'évaluation des entités nommées en français (FENEC : a balanced sample corpus for French named entity recognition). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, p. 82–94, Avignon, France : ATALA.

PASSONNEAU R. J., IDE N., SU S. & STUART J. (2014). Biber redux : Reconsidering dimensions of variation in American English. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics : Technical Papers*, p. 565–576, Dublin, Ireland : Dublin City University and Association for Computational Linguistics.

SEDDAH D. & CANDITO M. (2016). Hard time parsing questions : Building a QuestionBank for French. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 2366–2370, Portorož, Slovenia : European Language Resources Association (ELRA).

Actes des journées LIFT 2023
Nancy, France, 20-21 novembre 2023

The MeThAL Alsatian theater corpus and related resources: Work done and perspectives

Pablo Ruiz Fabo
Université de Strasbourg – Laboratoire LiLPa UR 1339, 67000 Strasbourg, France
ruizfabo@unistra.fr

ABSTRACT

As the MeThAL project nears completion, we present resources created and future perspectives. The first large TEI corpus of theater plays in Alsatian varieties was created (>500,000 words) and distributed under an open licence; a program for semi-automatic TEI-encoding based on OCR output was also released. A character prosopography in TEI was constructed; characters are annotated with metadata of potential socioliguistic relevance, like their gender or social class. An Alsatian emotion lexicon was developed, and a program to analyze the distribution of emotion terms in the plays, aggregating them by metadata like the characters' social variables or the plays's genre. Work related to Alsatian's large scriptolinguistic variation was performed, especially regarding emotion terms. To open the resources to a non-specialist public, a corpus exploration interface was created.

RÉSUMÉ

Le corpus MeThAL de théâtre en alsacien et ressources reliées : Réalisations et perspectives

Nous présentons les ressources développées et les perspectives, près de la clôture du projet MeThAL, qui a créé le premier grand corpus de théâtre alsacien en TEI (>500 000 mots), distribué sous licence ouverte. Un programme d'encodage TEI semi-automatique sur la base de sorties OCR a été distribué. Une prosopographie TEI de personnages a été construite ; elle décrit les personnages avec des variables sociales potentiellement pertinentes pour une analyse sociolinguistique (p. ex. genre des personnages, classe sociale). Un lexique d'émotions a été développé, ainsi qu'un programme permettant de l'appliquer pour analyser la distribution de ses termes dans le corpus. Le lexique est adapté à l'énorme variation scriptolinguistique des variétés alsaciennes. Pour ouvrir les ressources à un public non spécialisé, une interface d'exploration de corpus a été mise en ligne.

MOTS-CLÉS : théâtre alsacien, corpus TEI, analyse d'émotion, variation, sociolingustique.

KEYWORDS: Alsatian theater, TEI corpus, emotion analysis, variation, sociolinguistics.

CNRS, pages 113–118.

1 Introduction

The MeThAL project[1] has developed the first large public TEI-encoded corpus for Alsatian theater and related linguistic resources; the project's goals were presented at an early stage at LIFT's 2020 workshop. As the project nears completion, we would like to share with the same community the work done and perspectives.

2 Resources developed

In this section, we present the TEI-encoded corpus (2.1), the character prosopography (2.2), emotion analysis resources (2.3) and work on orthographic variation (2.4). Finally, we present the TEI-encoding workfow and corpus navigation interface (2.5).

2.1 TEI-encoded corpus

The first resource is the **TEI corpus**, with 77 plays (623,977 tokens), published in different locations and representing different varieties, by 27 authors. We included minor authors and well-known ones, both male and female, and several dramatic genres. Regarding genres, the corpus reflects comedy's predominance in the Alsatian tradition, but we also included dramas, popular dramas (*Volksstücke*) and tales (*Weihnachtsmärchen*). Out of the 77 plays, 51 are plays for which no previous electronic text existed; we carried out OCR based on image-mode digitizations at Strasbourg's national library (Bnu)[2] and a manual correction of the OCR prior to semi-automatic TEI encoding (section 2.5 describes the workflow). The remaining 26 plays had been published in wikitext format on Wikisource and, in those cases, our work consisted in converting the wikimarkup to TEI, using rules.[3]

As per FAIR practices, rich metadata were collected in the TEI header; an ODD with embedded schematron rules ensures metadata consistency. We assign a DOI to each resource via the Nakala data repository. The resources are exposed in a Nakala collection and via GitLab.[4]

[1] https://methal.pages.unistra.fr/en

[2] See https://www.numistral.fr/fr/theatre-alsacien

[3] The Wikisource transcriptions were carried out by wikipedian MireilleLibmann https://als.wikipedia.org/wiki/Text:August_Lustig/A._Lustig_S%C3%A4mtliche_Werke:_Band_2

[4] The Nakala collection is at https://nakala.fr/collection/10.34847/nkl.feb4r8j9. The GitLab repository is at https://git.unistra.fr/methal/methal-sources. Note that, as of this writing, the Nakala collection contains only 37 of the corpus plays. The rest are published on GitLab and are usable in their current version, having already been used for emotion analyses (Bernhard & Ruiz Fabo, 2022; Liu *et al.*, 2023; Ruiz Fabo *et al.*, 2024), but await more detailed verification before we assign a DOI to them via Nakala.

2.2 TEI feature structures for character social annotations

A **character prosopography in TEI** was also created (Ruiz Fabo *et al.*, 2024),[5] using the feature structures formalism (Romary, 2015), improving upon proposals for character description by Galleron (2017). Formalizing annotations with feature structures presents some degree of complexity, and consistency was ensured with the FS-VALIDATOR tool (Bermúdez Sabel, 2022), which generates ODD statements based on the feature structure declaration.

In the prosopography, characters are annotated with social variables such as their gender, social class, profession or professional category; we built a taxonomy of socioprofessional groups appropriate for the corpus context (Ruiz Fabo & Werner, 2021). These social variables have potential relevance for a sociolinguistic description of character speech. In section 3, we will discuss challenges related to analyzing such data.

2.3 Emotion analysis

A resource related to the corpus is Bernhard's ELAL (***Emotion Lexicon for Alsatian***);[6] it is based on French-Alsatian bilingual lexica and existing emotion lexica for French and German, aggregated into a large network. Emotion term variants are identified with graphical similarity methods. Thanks to this, the lexicon handles Alsatian's huge orthographic variability. The MeThAL corpus was used to extract emotion terms' variants (Bernhard & Ruiz Fabo, 2022), and the lexicon was then applied for first analysis of emotion vocabulary in Alsatian plays.

We applied the ELAL lexicon in EDYTHA (***Emotion Dynamics in Theater in Alsatian***), a program by Liu *et al.* (2023), which detects emotion vocabulary trends in a TEI corpus. It is inspired by Vishnubhotla & Mohammad's (2022) *TED* tool,[7] but improves its scoring via corpus-driven term-weighting schemes that prevent potential skews due to frequent lexicon terms. Taking advantage of the MeThAL character prosopography, EDYTHA aggregates emotion scores based on metadata like the character's gender, social class or professional group. This is interesting for purposes like assessing whether characters in a socially disfavoured position use more negative emotional terms than other groups. The tool also aggregates scores based on the plays' genre (e.g. comedies vs. dramas). The program was released under GPL.[8]

[5] https://git.unistra.fr/methal/methal-sources/-/tree/master/personography

[6] See https://nakala.fr/10.34847/nkl.40cex998 for the lexicon items and https://nakala.fr/10.34847/nkl.39b7617v for their emotion scores

[7] *TED* (Tweet Emotion Dynamics): https://github.com/Priya22/EmotionDynamics

[8] https://git.unistra.fr/methal/edytha

2.4 Approaches to orthographic variation

We also worked on Alsatian's large **orthographic variation**. Bernhard performed successful variant detection experiments, reported in Ruiz Fabo *et al.* (2024), using the ELAL lexicon as a training corpus with classical machine learning models and neural methods. Yang (2022) induced scriptural variation rules via methods in Millour (2020), applicable to multiple sequence alignment.

2.5 TEI-encoding automation and corpus navigation interface

A workflow for **semi-automatic TEI encoding** based on OCR outputs, using rules, a conditional random fields (CRF) model, and manual revision was described in Ruiz Fabo *et al.* (2024). We released the software and the CRF model under an open license (Briand & Ruiz Fabo, 2023). For OCR, we used Tesseract v4 (Smith, 2018), combining Fraktur and Latin script models, both language-specific and independent; we found that the tool and said models work well for the print sources from 1850 onwards used in the project. The CRF was implemented with `sklearn-crfsuite` (Korobov, 2019).

Finally, to engage the non-specialist public with this digitally underrepresented tradition, we developed a **navigation interface** that allows exploring the texts and metadata.[9]

3 Outlook: Towards sociolinguistic analyses?

The previous section shows that the corpus was useful for *computational literary studies* approaches, previously unattempted for Alsatian given lack of a corpus, but important because they add empirical variety to the field. Now the question arises whether sociolinguistic description could also benefit from the resources created.

The first caveat is of course that character speech is fictional and can only be socioliguistically relevant insofar as it reflects actual linguistic practice. Beyond that, in terms of geographic variation, it is unclear how to analyze the data: Do scriptural practices reflect the author's variety? The publisher's location? Fictional characters' intended varieties? Other types of variation may be more promising: Thanks to the prosopography metadata, we have over 28,000 speech turns annotated for character gender, 14,000 with the character's professional category and 11,000 with the social class.[10] These data may allow us to asses some variation trends in the future. The approach in Šeļa *et al.* (2023), which has succesfully identified distinctive features in character speech, sometimes tied to social factors, could be applied.

[9] The interface is at `https://methal.eu/ui/`.

[10] For these data and the way they were prepared, see the repository at `https://git.unistra.fr/methal/alsatian-character-speech`

Acknowledgements

This research was supported by Université de Strasbourg's IdEx program (Attractivité 2020 call).

I thank further project members, Delphine Bernhard, Dominique Huck, Pascale Erhart and Carol Werner for all collaborations, besides Alice Millour, who co-supervised one of the project internships.

We also thank our interns: Nathanaël Beiner, Andrew Briand, Lena Camillone, Hoda Chouaib, Audrey Deck, Barbara Hoff, Valentine Jung, Salomé Klein, Audrey Li-Thiao-Té, Qinyue Liu, Kévin Michoud, Alexia Schneider, Vedisha Toory, Heng Yang.

We also acknowledge the High Performance Computing Center of the University of Strasbourg for scientific support and access to computing resources. Part of the computing resources were funded by the Equipex Equip@Meso project (Programme Investissements d'Avenir) and the CPER Alsacalcul/Big Data.

References

BERMÚDEZ SABEL H. (2022). FS-Validator. https://github.com/HelenaSabel/FS-Validator.

BERNHARD D. & RUIZ FABO P. (2022). ELAL: An Emotion Lexicon for the Analysis of Alsatian Theatre Plays. In *Proceedings of LREC 2022*, p. 5001–5010, Marseille: ELRA.

BRIAND A. & RUIZ FABO P. (2023). FETE: Fast Encoding of theater in TEI. https://git.unistra.fr/methal/FETE.

GALLERON I. (2017). Conceptualisation of Theatrical Characters in the Digital Paradigm: Needs, Problems and Foreseen Solutions. *Human and Social Studies*, **6**(1), 88–108.

KOROBOV M. (2019). sklearn-crfsuite. https://github.com/TeamHG-Memex/sklearn-crfsuite.

LIU Q., RUIZ FABO P. & BERNHARD D. (2023). Towards emotion analysis for Alsatian theater. In *Computational Humanities Research (Posters)*. DOI : 10.5281/zenodo.8404253.

MILLOUR A. (2020). *Myriadisation de ressources linguistiques pour le traitement automatique de langues non standardisées*. PhD Thesis, Sorbonne Université. HAL : tel-03083213.

ROMARY L. (2015). Standards for language resources in ISO – Looking back at 13 fruitful years. *edition - Die Fachzeitschrift für Terminologie*, **11**(2), 13–19.

RUIZ FABO P., BERNHARD D., BRIAND A. & WERNER C. (2024). Computational drama analysis from almost zero electronic text. In M. ANDRESEN & N. REITER, Éds., *Computational Drama Analysis: Reflecting Methods and Implementations*. De Gruyter. https://univoak.eu/islandora/object/islandora%3A157880.

RUIZ FABO P. & WERNER C. (2021). Exploration du théâtre alsacien à travers ses listes de personnages pendant la période 1870-1940. In *Humanistica 2021*. DOI : 10.5281/zenodo.4762733.

SMITH R. M. D. (2018). Tesseract (v4.0). https://github.com/tesseract-ocr/tesseract.

VISHNUBHOTLA K. & MOHAMMAD S. M. (2022). Tweet Emotion Dynamics: Emotion Word Usage in Tweets from US and Canada. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 4162–4176, Marseille, France: European Language Resources Association.

YANG H. (2022). Détection de la variation graphique dans une langue non standardisée : le cas des dialectes alsaciens. Mémoire de master. https://dumas.ccsd.cnrs.fr/dumas-03794680.

ŠEĻA A., NAGY B., BYSZUK J., HERNÁNDEZ-LORENZO L., SZEMES B. & EDER M. (2023). From stage to page: language independent bootstrap measures of distinctiveness in fictional speech. DOI : 10.48550/arXiv.2301.05659.

Actes des journées LIFT 2023
Nancy, France, 20-21 novembre 2023

Transcription Automatique de l'arabe parlé à Tunis : Un pont vers l'analyse linguistique

Charles Vancaeyzeele,[1] Daphne Gonçalves Teixeira[1] Mohamed Malek Bahri[1]
(1) Laboratoire Ligérien de Linguistique (LLL), 10 rue de Tours, BP 46527 - 45072 ORLEANS CEDEX 2 (FRANCE)
charles.vancaeyzeele@univ-orleans.fr,
daphne.goncalves-teixeira@univ-orleans.fr,
mohamed-malek.bahri@univ-orleans.fr

Résumé

Cet article expose un projet de transcription du corpus de l'arabe parlé à Tunis, mené par des doctorants du Laboratoire Ligérien de Linguistique. La langue, peu documentée et sans orthographe standard, présente des défis majeurs en transcription manuelle en raison des variations vocaliques. Pour pallier les biais, une transcription automatique a été explorée, utilisant des modèles existants pour d'autres langues. L'objectif était de réduire la charge de travail et de documenter la langue tout en évitant une référence stricte à l'arabe standard. La méthodologie inclut trois étapes : transcription automatique avec Whisper AI, translittération en caractères latins (normes de Buckwalter), et harmonisation des caractères divergents. Cette approche facilite la comparaison avec les transcriptions manuelles et permettra d'établir une convention orthographique pour les voyelles, ainsi que d'identifier les processus phonologiques en synchronie grâce à un répertoire phonémique constitué. Ce projet contribuera à une meilleure compréhension de l'arabe parlé à Tunis.

Abstract

Automatic Transcription of Arabic Spoken in Tunis : A Bridge to Linguistic Analysis

This article presents a project for the transcription of the spoken Arabic corpus in Tunis, conducted by doctoral students at the Laboratoire Ligérien de Linguistique. The language, which is poorly documented and lacks a standard orthography, poses significant challenges for manual transcription due to vocalic variations. To mitigate biases, an automated transcription was explored using existing models for other languages. The main aim was to reduce the workload and document the language while avoiding a strict reference to standard Arabic. The methodology involves three stages : automated transcription using Whisper AI, transliteration into Latin characters (Buckwalter standards), and harmonization of divergent characters. This approach facilitates comparison with manual transcriptions and will enable the

establishment of an orthographic convention for vowels, as well as the identification of phonological processes in synchrony through the compilation of a phonemic inventory. This project will contribute to a better understanding of spoken Arabic in Tunis.

MOTS-CLÉS : corpus oraux, transcription automatique, analyse distributionnelle, Arabe Tunisien.

KEYWORDS : oral corpora, automatic transcription, distributional analysis, Tunisian Arabic.

1 Introduction

Dans ce papier, nous explicitons le protocole et les problématiques associées au projet de transcription (automatique et manuelle) d'un corpus d'arabe parlé à Tunis. Ce projet est mené par des doctorants issus de disciplines différentes (Sociolinguistique, Phonologie et Traitement Automatique des Langues) au sein du Laboratoire Ligérien de Linguistique.

2 Corpus d'arabe parlé à Tunis

L'idée présentée ici émane d'un projet de thèse ayant pour objectif de tester la faisabilité méthodologique de la chaîne de traitement des Enquêtes Sociolinguistiques à Orléans sur une langue peu documentée, l'arabe parlé à Tunis. La transcription des 43 heures d'entretiens récoltés à Tunis se fixe le même objectif que les transcriptions du corpus des ESLO : une transcription orthographique permettant de naviguer dans un texte aligné au signal et qui offre à tout usager du corpus la possibilité d'étudier plus en profondeur le phénomène qui l'intéresse.

Cela se complexifie dès lors que la langue qu'on transcrit n'est ni reconnue ni dotée d'une orthographe usuelle. Cela oblige à faire un choix aux conséquences épistémologiques parfois lourdes : faire une description morpho-phonologique du parler à transcrire dans l'optique de le doter d'une convention orthographique au risque de ne tester que la phase de transcription ou bien faire l'économie d'une description morphophonologique - aussi urgente soit-elle - et transcrire sa langue maternelle en aménageant un système orthographique existant mais non institutionnel (ARABIZI (Dichy *et al.*, 2019)) pour documenter l'usage réel de la langue par ses locuteurs.

2.1 Transcription du corpus

L'arabizi n'est pas la seule possibilité de translittération de l'alphabet arabe, il existe d'autres codifications plus rigoureuses et plus stables mais le choix d'utiliser l'arabizi est dicté par son ergonomie considérable par rapport aux autres normes existantes puisqu'il ne présente aucun signe spécial ou diacritique qui alourdit l'opération de transcription par un outil comme le clavier. Les autres normes de translittération, quant à elles, s'inspirent de l'arabe standard et contiennent des signes spéciaux, des diacritiques parfois encombrants quand il s'agit de transcrire un grand corpus. Il s'agit également d'une translittération en adéquation avec le corpus en question puisqu'elle émane des usagers des réseaux dont le besoin est de communiquer comme à l'oral, donc en arabe parlé. Cependant, comme la codification est véhiculée par les locuteurs eux-mêmes, elle présente quelques imprécisions et fluctuations graphiques au niveau des consonnes non représentées par l'alphabet latin qu'il convient d'harmoniser avant la transcription A comme dans le tableau suivant :

Lettres arabes	Caractères arabizi	Caractères normalisés en AT
ء	2 ou rien	2
خ	kh/5/7'	5
ذ	dh	th
ش	ch/sh	ch
ص	s / S / 9	S
ض	d / D / 9'	6
ط	t / T /6 :	T
ظ	z / Z / 6‘ :	6
ع	3	3
غ	gh / 3'	4
ق	q / 9 / 8	9
ح	7	7

TABLE 1 – Normalisation des consonnes arabes non représentées par l'alphabet latin

2.2 La question des voyelles

Le triangle vocalique de l'arabe tunisien ne semble pas faire consensus dans la littérature. On retrouve par exemple chez (Cohen, 1973) et (Marçais, 1977) un maintien de la triade classique /a/, /i/ et /u/ même si d'autres auteurs comme (Chekili, 1982) postulent l'existence de quatre phonèmes vocaliques brefs et quatre phonèmes vocaliques longs dont le [ ] postérieur et son équivalent long décrits par (Barkat, 2000) comme un timbre issu d'emprunts aux langues latines comme le français. Mzoughi reprend quant à elle le système à trois timbres vocaliques brefs

/i/, /a/, /u/ doublés de leur correspondantes longues.

Nous choisissons d'utiliser le système à trois voyelles comme hypothèse à tester dans la transcription A en orthographiant les phones vocaliques par les graphèmes vocaliques latins les plus proches des réalisations. Cependant, à cause du biais du locuteur natif et de la nature de l'outil utilisé pour transcrire, le clavier qui affiche un nombre limité de caractères, certaines réalisations différentes ont été confondues. La transcription A constitue donc un matériau expérimental de réflexion où se mêlent l'encodage phonologique et phonétique. Par conséquent, le phonème /a/ sera toujours écrit "a", le phonème /u/ sera tantôt écrit "ou" tantôt "o" et le phonème /i/ sera écrit "e" ou "i".

La transcription A n'a pas pour objectif d'être idéale ou de remplacer le signal. Elle avait pour premier objectif d'entrer dans le signal et les données afin de noter les principaux problèmes rencontrés dans le but de les traiter dans la transcription B qui sera, elle, phonologique.

phonème vocalique	**réalisations possibles**
/a:/	[æ:] [e:]
/i:/	[i:]
/u:/	[u:]
/a/	[æ] [a]
/i/	[ɪ] [ə]
/u/	[ʊ] [o] [u]

Transcription A	**Transcription B**	**Traduction**
r-rjaal	r-rjaal	les hommes
l-9diim	l-9diim	l'ancien
l-maktoub	l-maktuub	le destin
rja3-t	rja3-t	je suis revenu
3elm	3ilm	la science
9ol-t	9ul-t	j'ai dit

TABLE 2 – Exemples de Transcription des Réalisations Vocaliques

L'établissement de la transcription B qui contient les phonèmes de l'arabe tunisien a été établie à partir de l'instinct de locuteur natif mais elle présente des limites notamment au niveau de la longueur phonologique qui pourra être vérifiée à travers une analyse phonologique des données de la transcription A.

3 Transcription Automatisée

De l'ensemble du corpus, 20 heures ont été transcrites manuellement par un seul transcripteur natif avec un but exploratoire. Cette première démarche a permis d'identifier un nombre important de questions d'ordre linguistique, technique et méthodologique, à savoir, la diversité des réalisations (timbre) vocaliques, la différence de qualité des enregistrements et la normalisation pour transcrire la variation.

Au cours des dernières années, des initiatives ont été entreprises pour exploiter les avancées de la recherche en traitement automatique des langues (TAL) en vue de générer des transcriptions automatiques pour des langues peu documentées. Un exemple notable de cette initiative est le projet CREAM, qui se consacre au développement de technologies visant à documenter diverses langues créoles (Ferrand *et al.*, 2023). Ce projet a obtenu des résultats prometteurs en créant avec succès des transcriptions automatiques de données directement à partir du terrain.

Bien que le présent projet ne soit pas directement intégré au projet ANR CREAM, il s'inspire largement de l'initiative du projet CREAM visant à développer des outils pour la documentation de langues présentant un faible niveau de documentation. À l'instar des langues créoles, caractérisées par une base lexicale d'origine indo-européenne et une documentation rare, souvent non fondée sur une étude empirique des corpus, la variante de l'arabe parlée à Tunis partage des similitudes. Cette variante, tout comme les langues créoles, demeure insuffisamment décrite et peu explorée à travers des corpus de parole spontanée.

Dans le même esprit, nous avons choisi d'explorer les technologies existantes destinées aux langues largement documentées, en vue de générer des transcriptions automatiques pour une partie du corpus. Cette initiative vise plusieurs objectifs, notamment la comparaison des coûts entre une transcription manuelle et la création des outils nécessaires à la transcription automatique d'une telle langue. Un autre objectif essentiel est de tenter d'atténuer le biais de confirmation du transcripteur (natif) influencé par sa propre variation, qui peut potentiellement survenir lorsqu'une seule personne est chargée de la création d'un guide de transcription et de la transcription de l'intégralité du corpus.

En ce qui concerne la transcription automatique, la lacune apparente en matière de modélisation spécifique de l'arabe parlé à Tunis (et l'insuffisance des ressources pour en élaborer un à partir de zéro) pour la reconnaissance vocale requiert une approche d'adaptation à partir de modèles préexistants. L'objectif principal de cette tâche n'est pas de générer des transcriptions parfaites, mais plutôt de produire des transcriptions capables de réduire la charge de travail liée à la transcription manuelle, de rendre possible la documentation de cette langue tout en se méfiant d'une référence obligatoire à l'arabe standard et d'une approche intuitive.

4 Méthodologie de Travail

Pour obtenir ce type de transcription, nous avons entrepris une première évaluation en travaillant avec deux entretiens enregistrés, d'une durée d'environ une heure chacun, présentant des niveaux de qualité différents. Ces enregistrements ont été soumis à un processus en trois étapes pour générer les transcriptions. Dans la première phase de notre étude, nous utilisons les enregistrements audio bruts des entretiens en tant qu'entrée pour exécuter l'algorithme de transcription Whisper AI (Radford *et al.*, 2022), en privilégiant le modèle arabe large. Bien que nous ayons effectué des essais avec les modèles "medium" et "large", nous avons constaté une nette amélioration de la qualité lors du passage au modèle "large", ce qui a dicté notre choix. Cette première étape génère une transcription en graphie arabe qui servira de base dans la deuxième phase du processus où nous procéderons à la translittération de l'arabe vers l'alphabet latin.

La deuxième phase de notre processus, à savoir la translittération, repose sur l'utilisation du paquet "CAMeL Tools" (Obeid *et al.*, 2020) pour convertir la sortie de la première étape en une translittération correspondant aux normes de Buckwalter, qui définissent la correspondance entre les lettres arabes et les caractères latins. Il est toutefois important de noter que cette norme diffère de celle mise en place par le transcripteur pour effectuer la transcription manuelle, ce qui peut entraîner des difficultés lors de la comparaison des résultats. Afin de résoudre cette divergence, une troisième étape a été intégrée dans notre processus, au cours de laquelle la translittération générée est soumise à un ensemble de règles visant à harmoniser les caractères qui diffèrent de la norme de Buckwalter. Cette démarche garantit une plus grande cohérence et facilite la comparaison des données translittérées avec la transcription manuelle établie au préalable.

WhisperAI	اليوم أصباح مشيت للفاكولتي قريت وبعد روحت وبعد باش نمشي للتبيب
CAMel Tools	aliwm >SbaH m$it llfakwlti qrit wbEd rwHt wbEd ba*nm*i lltbib
Avec Règles	aliwm SSba7 mchit "llfakwlti" 9rit wmb3d rw7t wmb3d bach nmchi lltbib
Golden Standard	Lyoum S-Sbaa7 mchi-t li l-faculté 9arri-t w m-ba3d rawwa7-t w m-ba3d b-ech ne-mchi li l- b-ech ne-mchi li T-Tbiib

TABLE 3 – Exemple de sortie

Les résultats de cette chaîne de traitement sont récapitulés dans le tableau ci-dessus,

présentant les sorties de WhisperAI, CAMel Tools après l'application des règles, ainsi que le Golden Standard. Actuellement, des ajustements sont en cours pour améliorer le traitement des sorties des voyelles et obtenir des résultats encore plus précis.

5 Perspectives d'exploitation linguistique

À l'issue de ces trois étapes successives, nous obtenons en sortie un fichier contenant la transcription générée par l'outil de transcription Whisper AI (en graphie arabe), la translittération produite par CAMeL Tools, ainsi que la translittération adaptée. Cette approche permet une présentation plus claire des différentes phases du processus, facilitant ainsi la révision des résultats obtenus. De plus, cette structuration du fichier favorise la comparaison avec les transcriptions manuelles établies, contribuant ainsi à une évaluation plus précise de la qualité des transcriptions générées.

5.1 Analyse distributionnelle

C'est à partir de ce stade que la possibilité d'une exploitation linguistique des données dans leur ensemble se profile, notamment sur le plan phonologique. Plus précisément, il s'agit de comparer la transcription morpho-phonétique effectuée manuellement en amont avec la translittération morpho-phonémique adaptée de la sortie générée par CAMeL Tools, dans le but de vérifier la cohérence entre les graphèmes à valeur phonétique encodés dans la première et les graphèmes représentant les phonèmes, ou unités minimales distinctives, de l'arabe tunisien dans la seconde. Plus précisément, nous recourons à une analyse distributionnelle (une méthode propre à la théorie phonologique) pour justifier le rassemblement d'un ensemble de réalisations phoniques, e.g. [o], [u] et [ʊ], au sein d'un seul phonème, e.g. /u/.

L'analyse distributionnelle, originellement introduite dans (Troubetzkoy, 1949), consiste à observer la distribution (i.e. l'ensemble des contextes d'apparition) de chaque occurrence des phones recensés dans un corpus afin d'en déduire la relation d'opposition qu'ils entretiennent. Cette relation peut prendre deux formes différentes : si deux voire plusieurs phones s'opposent, ils sont associés à autant de phonèmes distincts ; s'ils ne s'opposent pas, ils sont associés à un seul et unique phonème. Une opposition ne peut être révélée que s'il est démontré que les phones concernés commutent (i.e. qu'ils induisent un changement de sens du mot auquel ils appartiennent dans le cas où ils sont interchangés). Il existe plusieurs types de distribution, chacune étant indicatrice d'une relation d'opposition spécifiée. Nous employons deux types de distribution, qui sont affichées ci-dessous. Soit deux phones x et y ainsi que deux contextes A et B :

— Distribution complémentaire : x apparaît exclusivement en A, et y apparaît exclusivement en B ; OU y apparaît exclusivement en A, et x apparaît exclusivement en

B

— Distribution libre : x et y peuvent apparaître en A comme en B

Une distribution complémentaire implique nécessairement que les phones concernés ne s'opposent pas, i.e. qu'ils sont associés à un seul et même phonème. En effet, puisqu'il est par définition impossible de les interchanger, il n'est pas possible qu'une alternance potentielle entre les deux induise de changement de sens d'un mot qui contiendrait l'un ou l'autre phone. En revanche, une distribution libre peut provoquer ou non un changement de sens dans un mot, et donc conditionner la présence comme l'absence d'une opposition.

5.2 Conclusions

Ce travail permettra d'extraire par exemple un ensemble de distribution pour chaque phone vocalique présent dans la troisième sortie, et de vérifier automatiquement la nature desdites distributions pour en déduire leur statut phonologique, d'après les principes de l'analyse distributionnelle. Le répertoire phonémique alors constitué, comparé aux répertoires proposés antérieurement dans la littérature, nous permettra de réaliser deux tâches. Premièrement, nous établirons une convention orthographique des voyelles à partir d'une correspondance systématique entre graphèmes et phonèmes. Deuxièmement, les contextes établis serviront à identifier divers processus phonologiques en synchronie.

Références

Barkat M. (2000). Détermination d'indices acoustiques robustes pour l'identification automatique des parlers arabes. Thèse de doctorat, Lyon 2.

Chekili F. (1982). The morphology of the Arabic dialect of Tunis.

Cohen D. (1973). "variantes, variétés dialectales et contacts linguistiques en domaine arabe". Bulletin de la Société Linguistique de Paris, **XVIII**(1), 215--248.

Dichy J., Driss F.-Z., Lagraf H. & Mostefa D. (2019). Communication écrite sur internet et par sms en arabizi : analyse de corpus relevant des glosses dialectales libanaise et marocaine. In A. Barontini, M.-A. Germanos, J. Guerrero, C. Pereira & C. Miller, Éds., AIDA2017, , Aix-en-Provence : IREMAM.

Ferrand E., Henri F., Lecouteux B. & Schang E. (2023). Application of speech processes for the documentation of kréyòl gwadloupéyen.

Marçais P. (1977). Esquisse grammaticale de l'arabe maghrébin. Paris : Librairie Adrien-maisonneuve.

Mzoughi I. (2015). Intégration des emprunts lexicaux au français en arabe dialectal tunisien. Thèse de doctorat.

Obeid O., Zalmout N., Khalifa S., Taji D., Oudah M., Alhafni B., Inoue G., Eryani F., Erdmann A. & Habash N. (2020). Camel tools : An open source

python toolkit for arabic natural language processing. In LREC2020, p. 7022--7032, Marseille, France : European Language Resources Association.

Radford A., Kim J., Xu T., Brockman G., McLeavey C. & Sutskever I. (2022). Robust speech recognition via large-scale weak supervision.

Roche M. (2010). Base, thème, radical. Recherches linguistiques de Vincennes, (39), 95--134.

Troubetzkoy N. S. (1949). Principes de Phonologie. Klincksieck, 2005 édition.