AfIA

Association française
pour l'Intelligence Artificielle

# TALN-RECITAL

*Conférence sur le Traitement Automatique des Langues Naturelles*

# PFIA 2019

# Table des matières

# Éditorial

La 26e édition de la conférence TALN et la 21e édition de la session jeunes chercheuses et chercheurs RECITAL se déroulent cette année à Toulouse au sein de la Plateforme française d'intelligence artificielle (PFIA). TALN a une longue tradition de tenue conjointe avec des conférences de domaines proches. Cette pratique a été initiée avec les Journées d'étude sur la parole (JEP) en 2002 à Nancy puis depuis 2008 tous les quatre ans (2008 : Avignon, 2012 : Grenoble, 2016 : Paris). Elle s'est diversifiée avec la Conférence de recherche d'information et applications (CORIA) en 2018 à Rennes. Elle innove cette année avec un hébergement à Toulouse au sein de PFIA. Ces événements sont l'occasion de rencontres enrichissantes pour tous. Cette année, ce ne sont pas moins de huit conférences, sans compter les ateliers associés, aux sessions desquelles les participants à TALN-RECITAL pourront se mêler : APIA (5e Conférence sur les Applications Pratiques de l'Intelligence Artificielle), CAp (21e Conférence sur l'Apprentissage Automatique), IC (30es Journées Francophones Ingénierie des Connaissances), JFPDA (14es Journées Planification, Décision et Apprentissage), JFSMA (27es Journées Francophones sur les Systèmes Multi-Agents), JIAF (13es Journées d'Intelligence Artificielle Fondamentale), RJCIA (17e Rencontre des Jeunes Chercheurs en Intelligence Artificielle), ainsi que CNIA (22e Conférence Nationale en Intelligence Artificielle), qui regroupe les thématiques de l'intelligence artificielle non couvertes par les conférences précédentes.

Les conférences invitées plénières, les sessions de présentations affichées et de démonstrations, les déjeuners et pauses café, les dîners de la conférence sont autant de moments programmés pour que se retrouvent les participants de toutes les conférences. Nous tenons à saluer la qualité de la planification et du suivi du comité scientifique de la plateforme ainsi que le grand travail du comité d'organisation, le tout visant à assurer que l'ensemble des conférences se tienne dans les meilleures conditions et au meilleur coût.

Pour la deuxième année consécutive, les modalités de soumission à TALN se faisaient avec un appel unique et un seul format de soumission en article court pouvant être étendu en article long sur proposition du comité de programme (et demande préalable des auteurs). Nous avons ainsi reçu soixante cinq articles courts et le comité de programme a proposé à dix articles le passage en format long (15 %) et a retenu trente et un articles en format court (48 %). Chaque article a été relu par trois membres du comité de lecture en s'appuyant le cas échant sur des relecteurs additionnels. Le comité de programme s'est appuyé sur ces relectures pour sélectionner lors d'une réunion pleinière les articles composant le programme. C'est un fonctionnement auquel nous sommes profondément attachés pour assurer une diversité dans les thématiques abordées. L'ensemble des évaluations ont été réalisées en double aveugle. Nous remercions les membres des comités de programme et de lecture (à parité femme – homme) pour leur contribution indispensable à ce processus. Le programme de la conférence est complété par quatre démonstrations sélectionnées par le comité de programme. Les titres des sessions donnent une idée des thématiques abordées par la conférence. Ils comprennent des paliers et tâches habituels du TAL (Morphologie et Syntaxe, Syntaxe, Résolution d'anaphores, Multilinguisme), reflètent la place prise par l'apprentissage (Apprentissage par transfert et modèles de langue, Plongements de mots), l'importance fondamentale que continuent à jouer les corpus et bases de données lexicales (Ressources), et l'intérêt du TAL pour des domaines particuliers (Langues spécialisées, Traitement de la langue biomédicale). Comme chaque année, l'ATALA a décerné un prix de thèse dont la récipiendaire présentera son travail en session plénière. La conférence a invité la présentation d'instruments récents du CNRS par leurs coordinatrices : d'une part le pré-GDR TAL (INS2I / informatique), qui adopte une vision inclusive du traitement de la langue (écrite, orale, signée), couvrant les communautés du traitement automatique des langues, du traitement automatique du langage parlé et de la recherche d'information ; d'autre part le GDR LIFT (INSHS / sciences du langage) sur la linguistique informatique, formelle et de terrain.

Cette année, dix-sept articles ont été soumis à RECITAL. Après avoir été chacun évalué par deux membres du comité de programme, quatre articles ont été retenus pour une présentation orale (soit un taux de sélection pour présentation orale de 24 %), et sept autres ont été retenus pour une présentation sous forme de poster (taux de sélection global de 65 %). Nous avons ainsi pu donner l'opportunité à douze jeunes chercheuses et chercheurs, en grande majorité en début de thèse, de présenter leurs travaux à la communauté. Nous remercions le comité de programme (également à parité femme – homme) pour leur minutieux travail de relecture.

Nous souhaitons pour finir au public de ces conférences une semaine riche en découvertes scientifiques et en rencontres de nouveaux collègues, dans une ambiance assurément chaude pour toute la semaine.

Emmanuel Morin, Sophie Rosset et Pierre Zweigenbaum (TALN)
Anne-Laure Ligozat et Sahar Ghannay (RECITAL)

# Comités

## Présidents de TALN

— Emmanuel Morin (LS2N, Université de Nantes)
— Sophie Rosset (LIMSI, CNRS, Université Paris-Saclay)
— Pierre Zweigenbaum (LIMSI, CNRS, Université Paris-Saclay)

## Membres du CP de TALN

— Delphine Bernard (LiLPa, Université de Strasbourg)
— Chloé Braud (LORIA, CNRS)
— Nathalie Camelin (LIUM, Le Mans Université)
— Peggy Cellier (IRISA, INSA Rennes)
— Benoît Crabbé (LLF, Université Paris Diderot)
— Iris Eshkol-Taravella (MoDyCo, Université Paris Nanterre)
— Cécile Fabre (CLLE-ERSS, Université Toulouse - Jean Jaurès)
— Núria Gala (LPL, Aix Marseille Université)
— Thierry Hamon (LIMSI, Université Paris Nord)
— Philippe Langlais (RALI/DIRO, Université de Montréal)
— Gwénolé Lecorvé (IRISA, Université de Rennes 1)
— Aurélie Névéol (LIMSI, CNRS, Université Paris-Saclay)
— Damien Nouvel (ERTIM, INaLCO)
— Didier Schwab (LIG, Université Grenoble Alpes)
— Xavier Tannier (LIMICS, Université Pierre et Marie Curie)

## Comité de lecture de TALN

— Gilles Adda (LIMSI, CNRS, Université Paris-Saclay)
— Salah Ait-Mokhtar (Naver Labs Europe)
— Alexandre Allauzen (LIMSI, CNRS, Université Paris-Saclay)
— Maxime Amblard (LORIA, Université de Lorraine)
— Jean-Yves Antoine (LIFAT, Université de Tours)
— Loïc Barrault (LIUM, Le Mans Université)
— Denis Béchet (LS2N, Université de Nantes)
— Frederic Béchet (LIS, Aix-Marseille Université)
— Patrice Bellot (LIS, Aix-Marseille Université)
— Asma Ben Abacha (Lister Hill Center, National Library of Medicine)
— Laurent Besacier (LIG, Université Grenoble Alpes)
— Yves Bestgen (ILC, Université catholique de Louvain)
— Philippe Blache (LPL, CNRS, Aix-Marseille Université)
— Fethi Bougares (LIUM, Le Mans Université)
— Thierry Charnois (LIPN, Université Paris 13)
— Vincent Claveau (IRISA, CNRS)
— Chloé Clavel (LTCI, Télécom ParisTech)
— Kevin Bretonnel Cohen (University of Colorado School of Medicine)
— Béatrice Daille (LS2N, Université de Nantes)
— Géraldine Damnati (Orange Labs)
— Gaël Dias (GREYC, Normandie Université)
— Marco Dinarelli (LIG, CNRS)
— Patrick Drouin (OLST, Université de Montréal)
— Dominique Estival (MARCS, Western Sydney University)
— Yannick Estève (LIUM, Le Mans Université)
— Olivier Ferret (CEA LIST)
— Karën Fort (STIH, Sorbonne Université)
— Thomas Francois (CENTAL, Université catholique de Louvain)
— Éric Gaussier (LIG, Université Grenoble Alpes)
— Jérôme Goulian (LIG, Université Grenoble Alpes)

— Natalia Grabar (STL, CNRS)
— Cyril Grouin (LIMSI, CNRS, Université Paris-Saclay)
— Olivier Hamon (Syllabs)
— Nabil Hathout (CLLE-ERSS, CNRS)
— Amir Hazem (LS2N, Université de Nantes)
— Nicolas Hernandez (LS2N, Université de Nantes)
— Stéphane Huet (LIA, Université d'Avignon et des Pays de Vaucluse)
— Christine Jacquin (LS2N, Université de Nantes)
— Sylvain Kahane (Modyco, Université Paris Nanterre)
— Olivier Kraif (LIDILEM, Université Grenoble Alpes)
— Mathieu Lafourcade (LIRMM, Université de Montpellier)
— David Langlois (LORIA, Université de Lorraine)
— Eric Laporte (LIGM, Université Paris-Est Marne-la-Vallée)
— Thomas Lavergne (LIMSI, Université Paris Sud, Université Paris-Saclay)
— Joseph Le Roux (LIPN, Université Paris 13)
— Benjamin Lecouteux (LIG, Université Grenoble Alpes)
— Yves Lepage (Waseda University)
— Denis Maurel (LIFAT, Université de Tours)
— Richard Moot (LIRMM, CNRS)
— Véronique Moriceau (IRIT, Université Paul Sabatier)
— Philippe Muller (IRIT, Université Paul Sabatier)
— Alexis Nasr (LIS, Aix Marseille Université)
— Adeline Nazarenko (LIPN, Université Paris 13)
— Luka Nerima (Université de Genève)
— Jian-Yun Nie (RALI/DIRO, Université de Montréal)
— Yannick Parmentier (LORIA, Université de Lorraine)
— Sebastian Peña Saldarriaga (Dictanova)
— Thierry Poibeau (Lattice, CNRS)
— Alain Polguère (ATILF, Université de Lorraine)
— Jean-Philippe Prost (LIRMM, Université de Montpellier)
— Solen Quiniou (LS2N, Université de Nantes)
— Christian Raymond (IRISA, INSA Rennes)
— Christian Retoré (LIRMM, Université de Montpellier)
— Djamé Seddah (ALMAnaCH, Paris Sorbonne Université)
— Gilles Serasset (LIG, Université Grenoble Alpes)
— Michel Simard (NRC, Canada)
— Kamel Smaili (LORIA, Université de Lorraine)
— Pascale Sébillot (IRISA, INSA Rennes)
— Ludovic Tanguy (CLLE-ERSS, Université Toulouse - Jean Jaurès)
— Juan-Manuel Torres-Moreno (LIA, Université d'Avignon et des Pays de Vaucluse)
— Guillaume Wisniewski (LIMSI, Université Paris-Sud, Université Paris-Saclay)
— François Yvon (LIMSI, CNRS, Université Paris-Saclay)

### Relecteurs additionnels de TALN

— Jingshu Liu (Dictanova)
— Emile Chapuis (LTCI, Télécom ParisTech)
— Caroline Langlet (LTCI, Paris Sorbonne Université)
— Joseph Lark (Dictanova)
— Alexandre Garcia (LTCI, Télécom ParisTech)

### Présidentes de RECITAL

— Anne-Laure Ligozat (LIMSI, CNRS, Université Paris-Saclay)
— Sahar Ghannay (LIMSI, CNRS, Université Paris-Saclay)

### Membres du CP de RECITAL

— Jean-Yves Antoine (LIFAT, Université de Tours)

— Ismail Badache (ESPE / LIS, Aix-Marseille Université)
— Amira Barhoumi (LIUM, Université du Maine - MIRACL Sfax)
— Rachel Bawden (University of Edinburgh)
— Aurélien Bossard (LIASD, Université Paris 8)
— Chloé Braud (LORIA, CNRS)
— Nathalie Camelin (LIUM, Université du Maine)
— Rémi Cardon (STL, Lille)
— Peggy Cellier (IRISA, INSA Rennes)
— Antoine Doucet (L3i, Université de la Rochelle)
— Maha Elbayad, LIG/ Inria
— Arnaud Ferré (LIMSI-CNRS/MaIAGE-INRA, Université Paris-Saclay)
— Amel Fraisse (Gériico, Lille)
— Thomas François (CENTAL, Université catholique de Louvain)
— Nicolas Hernandez (LS2N, Université de Nantes)
— Yann Mathet (Greyc, Université de Caen)
— Alice Millour (STIH, Université Paris-Sorbonne)
— Anne-Lyse Minard (LLL, Orléans)
— Jose Moreno (IRIT, UPS)
— Tsanta Randriatsitohaina (LIMSI, Université Paris-Sud, Université Paris-Saclay)
— Loïc Vial (LIG, Université Grenoble Alpes)

# Volume II : Articles courts

# Analyse faiblement supervisée de conversation en actes de dialogue

Kate Thompson[1] Nicholas Asher [1] Philippe Muller[2] Jeremy Auguste [3]
(1) IRIT, CNRS, (2) IRIT, Université de Toulouse, (3) LIS, Université Aix-Marseille
kate.thompson@irit.fr, nicholas.asher@irit.fr, philippe.muller@irit.fr,
jeremy.auguste@lis-lab.fr

RÉSUMÉ

Nous nous intéressons ici à l'analyse de conversation par *chat* dans un contexte orienté-tâche avec un conseiller technique s'adressant à un client, où l'objectif est d'étiqueter les énoncés en actes de dialogue, pour alimenter des analyses des conversations en aval. Nous proposons une méthode légèrement supervisée à partir d'heuristiques simples, de quelques annotations de développement, et une méthode d'ensemble sur ces règles qui sert à annoter automatiquement un corpus plus large de façon bruitée qui peut servir d'entrainement à un modèle supervisé. Nous comparons cette approche à une approche supervisée classique et montrons qu'elle atteint des résultats très proches, à un coût moindre et tout en étant plus facile à adapter à de nouvelles données.

ABSTRACT

**Weakly supervised dialog act analysis**

We are interested here in conversation analysis in the form of written *chats* in a task-oriented context between a customer and technical assistant. Our objective is to provide dialog act labels to each utterance, as a source of information for downstream tasks. We experimented with a weakly supervised approach based on heuristic rules, a few development annotations, and an ensemble method that is used to annotate more data in a noisy manner. This can be leveraged to train a more classical supervised approach. We compare this method to a classically supervised model and show that results are comparable, at a fraction of the annotating cost and arguably a better generalisation.

MOTS-CLÉS : Dialogue, chat, actes de dialogue, apprentissage faiblement supervisé.

KEYWORDS: dialog, chat, dialog act, weakly supervised learning.

# 1 Introduction

L'analyse de conversation est une perspective intéressante dans le cadre de la gestion de la relation client, notamment à cause de l'essor des plate-formes de conseil mettant en relation des clients avec des conseillers pouvant les aider à régler des problèmes techniques ou commerciaux. Une part importante de ces échanges a lieu par écrit dans le cadre d'échange par *chat*. Alors que l'analyse de corpus de conversation sous forme orale a une longue histoire, avec des corpus reconnus (Godfrey *et al.*, 1992; Carletta, 2007), les conversations sous forme écrite permis par les nouvelles formes de media sociaux (chat, forums, microblogging) ont fait l'objet de moins d'attention, même si cela commence à changer, par exemple les travaux sur les forums (Wang *et al.*, 2012), le chat, notamment en anglais (Asher *et al.*, 2016) et récemment en français (Damnati *et al.*, 2016).

L'analyse en actes de dialogue est un premier niveau d'analyse de la conversation qui permet de caractériser la fonction de communication d'un énoncé ou d'un tour de parole, en déterminant si un énoncé est une offre, une réponse, un retour sur un échange, ou si l'énoncé a une fonction sociale ou en rapport avec une tâche à réaliser. Ce niveau sert souvent de base à des analyses plus précises des échanges. Peu de travaux ont étudié la particularité des conversations par chat pour ce niveau d'analyse, à quelques exceptions près, comme (Perrotin *et al.*, 2018) pour des conversations client-conseillers. Dans tous les cas, les approches automatiques de l'annotation en actes de dialogue utilisent une supervision forte, qui nécessite une quantité de données non négligeables, le corpus Switchboard étant un exemple typique (Ji *et al.*, 2016; Kumar *et al.*, 2018). Ces modèles sont très dépendants du type d'interaction et des sujets de conversation, et *a fortiori* rien n'indique qu'ils se généralisent à de l'interaction écrite.

Nous proposons ici une méthode fondée sur une supervision indirecte, à partir de règles d'annotation superficielles, qui permet d'entraîner un modèle d'ensemble utilisé pour annoter automatiquement un corpus large. Ce dernier corpus sert alors à superviser l'entraînement d'un autre modèle, selon une méthodologie popularisée sous le nom de *data programming* par ses auteurs (Ratner *et al.*, 2016; Bach *et al.*, 2017). Nous évaluons ici l'apport de cette méthode en la comparant à une méthode plus classiquement supervisée, en montrant que quelques dizaines de règles assez simples, essentiellement lexicales, permette d'atteindre une exactitude de 80%, qui n'est atteinte de façon supervisée qu'avec plusieurs milliers de conversations, le meilleur modèle classiquement supervisé atteignant 84-86% sur des données comparables (Perrotin *et al.*, 2018). Dans cette étude et la notre la typologie des actes de dialogue employée correspond à une granularité intermédiaire avec 10 étiquettes différentes, décrites ci-dessous.

## 2 Un modèle de "programmation par les données"

Pour développer un modèle indirectement ou faiblement supervisé, nous suivons la méthode de *data programming* définie par (Ratner *et al.*, 2016; Bach *et al.*, 2017), qui consiste en les étapes suivantes :

- l'écriture manuelle d'heuristiques de catégorisation, partielles et potentiellement contradictoires, mais en cherchant une bonne couverture des données ;
- l'apprentissage d'un modèle génératif qui cherche à approximer la probabilité conjointe des catégories prédites et de l'exactitude et la couverture des règles écrites ; on peut voir cette étape comme une forme d'ensemble de classifieurs pondérés en fonction de leur accord mutuel ;
- à partir de ce modèle, l'annotation automatique "bruitée" des données (c'est-à-dire avec une distribution de probabilités sur toutes les catégories au lieu d'une étiquette unique) ;
- enfin, un modèle supervisé standard peut-être appris sur ces données annotées, en cherchant à coller à la distribution des étiquettes (avec une fonction de perte ajustée à ce cadre).

Ce modèle est conçu au départ comme un outil d'extraction d'informations, notamment de relations entre entités nommées. Il permet sur ces tâches d'obtenir des scores proches d'approches supervisées, sans nécessiter d'annotation de données d'entraînement, l'annotation étant réservée au développement des règles et à l'évaluation. Les auteurs ont aussi montré que des non-experts pouvaient développer des règles dans un temps équivalent à de l'annotation manuelle classique, avec une fiabilité égale, sinon meilleure (Ratner *et al.*, 2016).

Nous adaptons ici cette approche au cas de la classification de textes, où l'énoncé d'un acte de dialogue est impliqué dans une relation unaire (le type d'acte de dialogue).

| Acte | N | % | sup. % | Description |
|---|---|---|---|---|
| STA | 479 | 45.20 | 39.16 | affirmation/apport d'information |
| INQ | 149 | 14.10 | 19.21 | demande d'information |
| ACK | 113 | 10.70 | 6.62 | acquiescement |
| OPE | 100 | 9.50 | 4.07 | ouverture du dialogue |
| PPR | 63 | 5.90 | 15.18 | proposition de résolution du problème |
| CLO | 49 | 4.50 | 5.48 | clôture du dialogue |
| PRO | 37 | 3.50 | 5.73 | énoncé du problème |
| CLQ | 34 | 3.20 | 1.74 | question de clarification |
| TMP | 27 | 2.50 | 2.43 | mise en pause du dialogue |
| OTH | 8 | 0.76 | 0.38 | autre |

TABLE 1 – Distribution des actes de dialogues dans le jeu de développement utilisé pour définir les heuristiques, avec pourcentages comparés à la distribution dans les données d'entraînement de l'approche supervisée de (Perrotin *et al.*, 2018).

# 3 Données utilisées

Pour cette expérience nous utilisons les données récoltées dans le cadre du projet Datcha [1] avec l'opérateur téléphonique Orange, portant sur l'étude de relation clients par chat, et constitué de conversations écrites entre un télé-conseiller et un client cherchant à résoudre un problème, technique ou commercial.

Nous avons sélectionné un sous-ensemble de conversations, dont la majeure partie sert d'ensemble d'entrainement pour le modèle génératif, une petite partie sert d'ensemble de développement pour la mise au point des heuristiques de catégorisation, et une autre petite partie sert de données de tests pour évaluer l'approche et pouvoir la comparer aux alternatives. Seule les parties test et développement sont manuellement annotées.

Les dialogues sont segmentés automatiquement, à partir des journaux d'interaction client-conseiller, qui liste les échanges verbaux et des métadonnées sur le contexte (service contacté, enquête de satisfaction par exemple) qui ne sont pas utilisées ici, à l'exception de certaines interventions automatique de la plate-forme de support, qui sont reliées au contexte de la conversation. Pour cela, chaque retour à la ligne d'un participant au chat est considéré comme délimitant la fin d'un acte, et nous appliquons le segmenteur en phrases de l'outil CoreNLP (Manning *et al.*, 2014) à chaque ligne, en plus de quelques heuristiques pertinentes pour des dialogues : segmentation sur les " ?", les ouvertures ou marques fréquentes d'ouverture d'actes de dialogue : *ok, merci, d'accord, bonjour, sinon*. Chaque acte ainsi délimité est censé ne correspondre qu'à un seul des types d'actes prévus.

La partie d'entrainement est constitué de 3000 conversations, le développement de 13 conversations segmentées et annotées en actes de dialogue, et le test de 2 conversations, ce qui correspond respectivement à 155k, 1059 et 181 segments.

Nous avons suivi le schéma d'annotation choisi par (Perrotin *et al.*, 2018) et appliqué sur des données similaires. La table 1 montre la distribution des types d'actes de dialogue dans les données de développement, et la comparaison avec la distribution reportée sur le corpus de (Perrotin *et al.*, 2018).

1. http://datcha.lif.univ-mrs.fr/

| Acte | Locuteur | énoncé |
|---|---|---|
| OPE | INFO | Vous entrez en conversation avec TC1. |
| INQ | TC | Que puis-je faire pour vous ? |
| CLQ | TC | sans exception ? |
| ACK | TC | Rassurez vous ,nous allons voir cela ensemble |
| PRO | CL | j'ai changé de forfait hier, j'ai pris le ... |
| PPR | TC | Je vous propose de recevoir par voie postale... |
| TMP | TC | je vous prie de rester en ligne |
| STA | CL | pas de tonalité |
| CLO | CL | bonne journée à vous |
| OTHER | CL | répondez moi svp |

TABLE 2 – Exemple d'énoncés d'actes de dialogue avec leur type pris dans différentes conversations. TC = téléconseiller, CL = client, INFO=intervention automatique de la plate-forme de support.

On peut noter des différences importantes sur les acquiescements et les ouvertures, probablement car le travail cité ne segmente pas à l'intérieur des tours de parole, quitte à donner un acte de dialogue "principal" pour le tour, ce qui peut faire négliger les catégories d'actes plus sociales que liées à la tâche (et représente une perte d'information que nous avons voulu éviter). De même il y a une grosse différence sur les actes de propositions de résolution du problème, qu'il est difficile d'expliquer autrement que par un biais d'échantillonnage. La table 2 montre un exemple de dialogue (extrait) avec les actes associés aux énoncés, et la table 2 montre des exemples d'énoncés pour chaque acte.

# 4 Expérimentations

Pour évaluer l'intérêt et les performances de l'approche légèrement supervisée, nous détaillons ici l'expérimentation faite en comparant un modèle "génératif" à partir des règles produites et des dialogues non annotés, ainsi qu'un modèle discriminatif entraîné sur ces données bruitées résultantes en section 4.1, et un modèle classiquement supervisé, section 4.2.

## 4.1 Modèle génératif à base de règles

Le modèle génératif repose sur un ensemble d'heuristiques (51) prédisant la classe des énoncés sur la base d'informations superficielles :

— patrons lexicaux spécifiques ;
— type du locuteur (conseiller ou client) ;
— position de l'énoncé dans le dialogue (proche du début/de la fin) ;
— contenu du contexte dialogique (type du locuteur des tours précédents et/ou suivants) ;

Ces règles ont été développées à partir d'un petit ensemble de dialogues annotés manuellement, sur lesquels leur couverture et précision sont estimées.

Ces règles sont par exemple de la forme :

— **si** le locuteur est le téléconseiller, et le tour commence (à n caractères près) par j (...) (vais|suis en train|vien|confirm| fai| prend|consult) et le tour ne contient pas " ?" **alors** le type d'acte est PPR (proposition de résolution de problème)
— **si** le locuteur est le client et le tour n'est pas social/une ouverture* et le tour précédent était par le télé-conseiller et contenait une forme de proposition d'aide* **alors** le type de l'acte est un PRO (énoncé du problème). Ici les parties de règle signalées avec un * sont exprimées sous forme d'expressions régulières sur la forme du tour, en listant les alternatives possibles (comme pour la règle précédente).

Une règle par défaut est déclenchée si aucune autre ne couvre le cas considérée, et attribue l'étiquette STA (affirmation), qui est la classe majoritaire.

Ces règles sont partielles, dans le sens où elles peuvent s'abstenir d'une décision sur une instance, et peuvent couvrir en partie les mêmes instances, de façon contradictoire ou non. Sur la base de ces règles et de données (non annotées) le modèle génératif peut donner la distribution jointe des classes et de la précision des règles par maximum de vraisemblance, cf (Ratner *et al.*, 2016) p. 4. En utilisant ces distributions sur les données de la partie d'entraînement (3000 dialogues), on peut alors entrainer un modèle supervisé de façon "bruitée". Nous utilisons l'implémentation Snorkel pour l'entrainement du modèle génératif [2], et un réseau de neurones récurrent pour apprendre la catégorisation des énoncés, en s'assurant de définir une fonction de perte qui prend en compte une distribution de probabilités comme référence au lieu d'un label unique (ici une mesure d'entropie croisée). Nous utilisons un bi-LSTM simple à une couche, qui utilise les 2 états finaux de la séquence comme entrée d'une couche finale avant un softmax sur les scores des classes. L'état du LSTM a une dimension de 256 les mots en entrée étant plongés dans un espace de 100 dimensions, initialisés avec des embeddings fastText (Bojanowski *et al.*, 2017) calculés sur le corpus d'entrainement, car ils donnent une certaine robustesse à ces représentations (indispensable vue la nature du langage utilisé dans les données).

Les réglages du LSTM ont été faits en observant les résultats sur le corpus de développement, sans faire des réglages très fins dans la mesure où les règles de départ sont déjà très biaisées par rapport au jeu de développement. Nous n'avons pas essayé non plus de reproduire exactement le modèle supervisé utilisé pour la comparaison dans la section suivante, calqué sur (Perrotin *et al.*, 2018), dans la mesure où nous ne pouvons savoir si les valeurs optimales pour une configuration (supervisée/générative) se généralise à l'autre. Tout juste sommes nous restés dans des espaces de paramètres et des capacités relativement comparables.

## 4.2 Modèle supervisé de comparaison

Le modèle supervisé utilisé dans nos expérimentations est le réseau de neurones décrit dans (Perrotin *et al.*, 2018), sans la couche CRF pour garder un modèle similaire à la partie précédente. C'est un réseau de neurones récurrent hiérarchique à deux niveaux. Le premier niveau permet de s'intéresser aux tours de paroles en prenant en entrée la séquence de mots de chaque tour. Le second niveau permet de prendre en compte l'ensemble de la conversation à partir des états cachés en sortie du premier niveau représentant les tours de paroles. Les deux niveaux sont des réseaux récurrents bidirectionnels de type LSTM. La couche de décision utilise les états cachés du LSTM du deuxième niveau afin d'obtenir une prédiction de l'acte de dialogue de chaque tour de parole. Les états cachés du premier

2. `github.com/HazyResearch/snorkel`

niveau sont de taille 64 et ceux du second niveau sont de taille 128. En entrée du premier niveau, les mots sont représentés par des embeddings de dimension 100 qui sont entraînés en même temps que le reste du réseau. L'information sur le scripteur de chaque tour est également donné en entrée du second niveau sous forme d'embeddings de dimension 5 entraînés par le réseau. L'entropie croisée est utilisée pour la fonction de coût et Adadelta est utilisé pour la rétropropagation du gradient.

Pour l'entraînement, nous utilisons les mêmes données que dans (Perrotin *et al.*, 2018). Dans ce corpus annoté manuellement, un seul acte de dialogue est attribué à chaque tour de parole et aucune segmentation supplémentaire n'est réalisée. Le corpus d'entraînement est composé de 2390 dialogues. Afin de pouvoir comparer ce modèle avec le modèle de la section précédente, nous utilisons ce modèle sur les 2 conversations de la partie de test.

## 4.3 Résultats et analyse

Le modèle génératif entraîné permet d'estimer l'apport des règles pondérées par le modèle et selon leurs paramètres estimés : nous reportons en table 3 les couvertures min, max et en moyenne des règles, ainsi que leur exactitude. Pour les modèles supervisés, le modèle entrainé sur les données

| Mesure | Min | Max | Moyenne |
|---|---|---|---|
| Exactitude | 0.4918 | 0.5035 | 0.4971 |
| Couverture | 0.6727 | 0.6846 | 0.6786 |

TABLE 3 – Performances des règles sur les données de test : exactitude du label prédit, couverture des instances par chaque règle.

annotées automatiquement par le modèle génératif atteint une exactitude de 80.1% sur le test, alors que le modèle supervisé classique n'atteint que 67.4% sur ces données de test. Il faut prendre ce score avec prudence car le modèle est entrainé sur une segmentation en actes différente, avec une distribution des étiquettes différentes comme on l'a vu. Evalué sur un jeu de test différent mais avec la même segmentation, (Perrotin *et al.*, 2018) reporte un score de 84% avec le même réseau de neurones, et 86% quand le réseau est adjoint à un CRF séquentiel. On peut noter que d'après cet article, il faut déjà 500 dialogues [3] annotés pour atteindre les 80% de la méthode non supervisée (même si là encore les conditions ne sont pas exactement les mêmes, l'ordre de grandeur est plausible).

En complément d'analyse, nous présentons les résultats par type d'acte de dialogue selon la méthode discriminante entrainée sur les données "automatiques" ou avec la supervision "classique" en table 4. Au vu de la petite taille du corpus de test, il est difficile de tirer des conclusions trop rapides sur certains labels dont le support est très bas et implique une variance importante. On peut simplement noter que les quatre classes les plus présentes dans le test sont les mêmes que dans le développement (avec un ordre légèrement différent), et que la généralisation est meilleure pour le modèle faiblement supervisé, alors que le modèle non supervisé est entrainé sur une distribution différente. Ceci peut expliquer l'écart sur ces données de test, mais il reste que le niveau de performance du modèle faiblement supervisé est comparable au supervisé quand chacun est "entrainé" sur une distribution comparable à son propre jeu de test.

3. Soit entre 20k et 30k instances d'actes.

| Type d'acte | P.S. | R.S. | F.S. | P.D. | R.D. | F.D. | Support |
|---|---|---|---|---|---|---|---|
| STA | 0.58 | 0.80 | 0.67 | 0.64 | 0.89 | 0.74 | 54 |
| OPE | 1.00 | 0.48 | 0.65 | 0.96 | 1.00 | 0.98 | 23 |
| PRO | 0.31 | 0.57 | 0.40 | 0.00 | 0.00 | 0.00 | 7 |
| INQ | 0.85 | 0.89 | 0.87 | 0.85 | 0.89 | 0.87 | 19 |
| CLQ | 0.62 | 0.62 | 0.62 | 1.00 | 0.50 | 0.67 | 8 |
| ACK | 1.00 | 0.50 | 0.67 | 0.89 | 0.71 | 0.79 | 34 |
| TMP | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 13 |
| PPR | 0.33 | 0.55 | 0.41 | 1.00 | 0.73 | 0.84 | 11 |
| CLO | 0.86 | 0.67 | 0.75 | 0.80 | 0.89 | 0.84 | 9 |
| OTHER | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 3 |
| Moyenne | 0.75 | 0.67 | 0.68 | 0.79 | 0.80 | 0.78 | 181 |

TABLE 4 – Résultats par type d'actes de dialogue, avec P(récision), R(appel), F(score) pour l'approche discriminante bruitée (D) et l'approche supervisée classique (S). Le support indique le nombre d'instances par classe.

# 5 Perspectives et conclusion

Nous avons présenté ici un modèle d'étiquettage en actes de dialogue légèrement supervisé, à partir d'heuristiques simples, de quelques annotations de développement, et une méthode d'ensemble sur ces règles qui sert à annoter automatiquement un corpus plus large de façon bruitée. Nous avons montré que ce modèle est compétitif avec une approche supervisée. Un autre avantage plausible de cette approche est de faciliter le transfert vers des données différentes, d'une part parce que la conception des règles est plus robuste, d'autre part parce qu'ajouter des spécificités de nouvelles données se fait aisément en adaptant les règles, quand un modèle supervisé nécessite soit de nouvelles annotations, soit une approche d'apprentissage par transfert aux résultats incertains. Cette hypothèse pourrait être testée avec des données nouvelles, ou bien déjà en séparant les données en sous-domaine selon les catégories de problème technique définies par l'opérateur sur sa plate-forme.

# Remerciements

Ce travail a été financé par l'Agence Nationale pour la Recherche, dans le cadre du projet DATCHA (ANR-15-CE23-0003).

# Références

ASHER N., HUNTER J., MOREY M., BENAMARA F. & AFANTENOS S. D. (2016). Discourse structure and dialogue acts in multiparty dialogue : the stac corpus. In *LREC*.

BACH S. H., HE B. D., RATNER A. & RÉ C. (2017). Learning the structure of generative models without labeled data. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, p. 273–282.

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

CARLETTA J. (2007). Unleashing the killer corpus : experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation*, **41**(2), 181–190.

DAMNATI G., GUERRAZ A. & CHARLET D. (2016). Web chat conversations from contact centers : a descriptive study. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*

GODFREY J. J., HOLLIMAN E. C. & MCDANIEL J. (1992). Switchboard : telephone speech corpus for research and development. In *[Proceedings] ICASSP-92 : 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, p. 517–520 vol.1.

JI Y., HAFFARI G. & EISENSTEIN J. (2016). A latent variable recurrent neural network for discourse-driven language models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 332–342, San Diego, California : Association for Computational Linguistics.

KUMAR H., AGARWAL A., DASGUPTA R. & JOSHI S. (2018). Dialogue act sequence labeling using hierarchical encoder with CRF. In *AAAI*, p. 3440–3447 : AAAI Press.

MANNING C. D., SURDEANU M., BAUER J., FINKEL J., BETHARD S. J. & MCCLOSKY D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, p. 55–60.

PERROTIN R., NASR A. & AUGUSTE J. (2018). Dialog Acts Annotations for Online Chats. In *25e conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Rennes, France.

RATNER A. J., SA C. D., WU S., SELSAM D. & RÉ C. (2016). Data programming : Creating large training sets, quickly. In *Advances in Neural Information Processing Systems 29 : Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, p. 3567–3575.

WANG L., KIM S. N. & BALDWIN T. (2012). The utility of discourse structure in identifying resolved threads in technical user forums. In *COLING*, p. 2739–2756 : Indian Institute of Technology Bombay.

# Apport de l'adaptation automatique des modèles de langage pour la reconnaissance de la parole: évaluation qualitative extrinsèque dans un contexte de traitement de cours magistraux

Salima Mdhaffar[1] Yannick Estève[2] Nicolas Hernandez[3] Antoine Laurent[1]
Solen Quiniou[3]
(1) LIUM, Avenue Olivier Messiaen ,72085 Cedex 9 Le Mans , France
(2) LIA, AGROPARC, BP 1228, 84911 Cedex 9 Avignon, France
(3) LS2N, 2 rue de la houssinière, BP 92208, 44322 Cedex 3 Nantes, France
firstname.lastname@{univ-lemans,univ-avignon,univ-nantes}.fr

## RÉSUMÉ

Malgré les faiblesses connues de cette métrique, les performances de différents systèmes de reconnaissance automatique de la parole sont généralement comparées à l'aide du taux d'erreur sur les mots. Les transcriptions automatiques de ces systèmes sont de plus en plus exploitables et utilisées dans des systèmes complexes de traitement automatique du langage naturel, par exemple pour la traduction automatique, l'indexation, la recherche documentaire... Des études récentes ont proposé des métriques permettant de comparer la qualité des transcriptions automatiques de différents systèmes en fonction de la tâche visée. Dans cette étude nous souhaitons mesurer, qualitativement, l'apport de l'adaptation automatique des modèles de langage au domaine visé par un cours magistral. Les transcriptions du discours de l'enseignant peuvent servir de support à la navigation dans le document vidéo du cours magistral ou permettre l'enrichissement de son contenu pédagogique. C'est à-travers le prisme de ces deux tâches que nous évaluons l'apport de l'adaptation du modèle de langage. Les expériences ont été menées sur un corpus de cours magistraux et montrent combien le taux d'erreur sur les mots est une métrique insuffisante qui masque les apports effectifs de l'adaptation des modèles de langage.

## ABSTRACT

**Contribution of automatic adaptation of language models for speech recognition : extrinsic qualitative evaluation in a context of educational courses**

Despite the known weaknesses of this metric, the performance of various automatic speech recognition systems is generally compared using the word error rate. The automatic transcriptions of these systems are usually used in complex natural language processing systems, for example for machine translation, indexation, document retrieval... Recent studies have proposed metrics to compare the quality of automatic transcriptions of different systems according to the target task. In this study, we investigated to qualitatively measure the contribution of the automatic adaptation of language models to the domain of a lecture. The transcriptions of the teacher's speech can serve as a basis for navigating in the video of the lecture or or for allowing the enrichment of its pedagogical content. By taking these two tasks into account, we propose to evaluate the contribution of the language model adaptation. Experiments were conducted on an educational corpus, and show how the word error rate is an insufficient metric that masks the strength contributions of the adaptation of language models.

MOTS-CLÉS : reconnaissance automatique de la parole, adaptation de modèles de langage, mesure

d'indexibilité, recherche d'information, éducation.

KEYWORDS: Automatic Speech Recognition, Language Model Adaptation, Word Error Rate, Indexability, Information Retrieval, Transcription, Educational Applications.

# 1 Introduction

La transcription automatique de cours magistraux convertit automatiquement le discours de l'enseignant (audio) en texte. Même si ces dernières années la technologie de reconnaissance automatique de la parole a considérablement progressé, principalement grâce aux architectures neuronales pour la modélisation acoustique, un système de reconnaissance automatique de la parole (SRAP) reste sensible aux mots hors vocabulaire et à la précision de ses modèles de langage (ML). Un tel système doit par exemple être bien préparé pour traiter des documents spécialisés. Or, dans le cadre de la transcription de cours magistraux, chaque cours nécessite une terminologie précise liée à son domaine. L'adaptation des modèle de langage est une technique indispensable pour résoudre ce problème d'inadéquation entre les données d'apprentissage et de test.

Généralement, pour la reconnaissance de la parole, le gain en performance de l'adaptation des modèles de langage est mesurée à l'aide du taux d'erreur mots (WER) (Pallett, 2003) : cette métrique d'évaluation est couramment utilisée dans la littérature pour l'analyse des performances des systèmes de reconnaissance automatique de la parole. Cette mesure s'appuie sur une comparaison entre la phrase produite par le SRAP et la phrase correspondante transcrite manuellement. Un alignement mot à mot utilisant la distance de Levenshtein est réalisé entre la transcription manuelle (référence) et la transcription automatique (hypothèse). Ensuite, une comparaison est effectuée selon les différents types d'erreurs sur les mots que peut commettre le système : insertions, suppressions et substitutions. Le calcul du WER s'effectue selon la formule suivante :

$$WER = \frac{S + I + D}{N} \tag{1}$$

où S est le nombre de mots substitués par le système, I est le nombre de mots insérés par le système, D est le nombre de mots supprimés par le système et N est le nombre total de mots dans la phrase.

WER attribue un score d'erreur en pourcentage pour la transcription globale. Cela est très utile lorsqu'il s'agit d'évaluer la performance du SRAP isolément. Cependant, les systèmes SRAP sont souvent conçus comme une brique dans d'autres applications de traitement de langage naturel qui utilisent les transcriptions de sortie pour effectuer d'autres tâches. Ces transcriptions constituent en effet une ressource précieuse pour d'autres modules technologiques appliquant des traitements tels que la recherche d'informations, la traduction, l'indexation de documents... La qualité des transcriptions de sortie affecte ainsi directement les performances de ces modules.

L'adaptation des MLs pour des cours magistraux a suscité beaucoup d'attention dans la littérature (Cerva *et al.*, 2012; Bell *et al.*, 2013; Yamazaki *et al.*, 2007; Kawahara *et al.*, 2008; Martínez-Villaronga *et al.*, 2013). La performance de ces travaux a été évaluée en utilisant WER ou la perplexité. Cependant, cette mesure ne prend pas en compte la gravité de l'erreur en fonction de la tâche finale (Luzzati *et al.*, 2014). En 2002, les auteurs de (Hürst *et al.*, 2002) ont déjà démontré que l'utilisation d'un vocabulaire thématique améliorait la reconnaissance de la parole et l'indexation des cours magistraux. Dans leurs analyses, ils se sont intéressés à l'analyse de l'impact du locuteur et du vocabulaire et ils ont démontré que l'utilisation d'un vocabulaire du domaine du cours améliorait les

performances du SRAP.

Considérant que le taux d'erreur de mot (WER) n'est pas suffisamment pertinent pour comparer la performance du SRAP pour certaines tâches spécifiques (Jannet *et al.*, 2015; Favre *et al.*, 2013) et que les erreurs du SRAP peuvent avoir une forte incidence sur la précision de plusieurs tâches de TALN, nous explorons, dans cet article, l'utilisation de deux mesures d'évaluation plus pertinentes pour comparer les apports de l'adaptation du modèle de langage pour un SRAP.

# 2 Données expérimentales

## 2.1 Description des données

Les expériences ont étés effectuées sur des données du domaine de l'éducation : elles sont constituées d'environ 10 heures d'enregistrements de cours magistraux en français. Toutes ces données ont été transcrites manuellement et annotées en segments thématiques par des experts. Les annotations ont été réalisées conformément au guide d'annotation développé durant le projet PASTEL (Mdhaffar *et al.*, 2018).
Ce corpus contient (1) des vidéos filmées durant des séances de cours magistraux, (2) les transcriptions manuelles de ces cours magistraux, (3) les diapositives de ces cours et (4) des annotations de segmentation thématique avec deux niveaux de granularité. « Granularité 1 » représente une segmentation fine du cours : chaque nouveau concept abordé durant le cours constitue un segment. « Granularité 2 » regroupe les segments de type « Granularité 1 » : elle est utilisée lorsqu'il y a un changement de sous-thème plus général qui permettrait d'arrêter l'apprentissage humain à ce moment-là et de reprendre plus tard l'apprentissage d'autres concepts. Le tableau 1 présente quelques statistiques du corpus (les cours pour lesquelles on dispose des diapositives). La deuxième et la troisième colonnes du tableau représentent, respectivement, le nombre de segments de « Granularité 1 » (G1) et le nombre de segments de « Granularité 2 » (G2). Le Nombre de locuteurs du corpus présenté dans le tableau est égale à 4.

## 2.2 Annotation en mots clés

Les mots du domaine ont été extraits manuellement à partir des transcriptions manuelles et des diapositives des cours. Nous considérons, comme mots du domaine, les expressions linguistiques faisant référence à des concepts, des objets ou des entités essentielles pour la compréhension de la diapositive actuelle ou d'une transcription donnée. Nous avons inclus tous les termes scientifiques et techniques ainsi que les acronymes et expressions permettant d'aller plus loin dans le sujet du cours. Ces annotations ont été réalisées pour les cours pour lesquels nous disposons des diapositives (6 cours du corpus). Les colonnes 4 et 5 du tableau 1 représentent, respectivement, le nombre de mots-clés annotés pour la transcription (Kw_$t$) et pour les diapositives (Kw_$s$). La dernière colonne contient la durée de chaque cours.

# 3 Description du SRAP et de l'adaptation du ML

## 3.1 SRAP

Le système SRAP est basé sur la boîte à outils Kaldi (Povey *et al.*, 2011). Des modèles acoustiques de type chain-TDNN (Povey *et al.*, 2016) ont été entraînés sur environ 300 heures de parole, princi-

| **Cours** | **G1** | **G2** | **Kw_$t$** | **Kw_$s$** | **Durée** |
|---|---|---|---|---|---|
| *Introduction à l'informatique* | 31 | 2 | 65 | 59 | 1h 04m |
| *Introduction à l'algorithmique* | 38 | 10 | 30 | 37 | 1h 17m |
| *Les fonctions* | 35 | 3 | 121 | 79 | 1h 14m |
| *Réseaux sociaux et graphes* | 43 | 7 | 74 | 97 | 1h 05m |
| *Algorithmique distribuée* | 72 | 5 | 314 | 158 | 1h 16m |
| *Langage naturel* | 52 | 5 | 130 | 107 | 1h 09m |
| *Total* | 271 | 32 | 734 | 537 | 7h 05m |

TABLE 1 – Statistiques du corpus

palement de type journaux d'actualités radio- ou télé-diffusés en français. Les modèles de langage génériques (n-grammes) ont été entraînés sur les transcriptions manuelles de la parole utilisée pour l'apprentissage des modèles acoustiques mais également sur des articles de journaux, pour un total de 1,6 milliard de mots. Le vocabulaire du modèle de langage générique contient environ 160 000 mots. Les détails sur les modèles de langage peuvent être trouvés dans (Rousseau *et al.*, 2014).

### 3.2 Adaptation du ML

Dans cette étude, nous émettons l'hypothèse que l'enseignant collabore *a minima* en fournissant les diapositives de son cours. Les titres des diapositives sont importants pour donner aux étudiants une idée rapide du contenu des parties d'un cours. Ils représentent ainsi souvent l'information principale sur laquelle l'étudiant s'appuie pour chercher et naviguer dans le cours. L'idée est donc d'utiliser les titres des diapositives comme des requêtes. En se basant sur le travail de (Lecorvé *et al.*, 2008), les requêtes sont soumises à un moteur de recherche (Google) et les pages pointées par les liens renvoyés sont téléchargées. Nous avons limité la recherche à 400 pages Web pour chaque requête. Le contenu textuel principal de ces pages est extrait pour construire un modèle de langage du domaine. Une adaptation du modèle de langage est faite par interpolation linéaire de deux modèles : le modèle générique et le modèle du domaine. Les mots les plus fréquents parmi les 400 pages Web récupérées sont ajoutés au vocabulaire du SRAP pour le traitement d'un cours magistral particulier.

## 4 Évaluation du ML

### 4.1 Méthodologie d'évaluation : tâche de recherche d'information

L'une des tâches qui peuvent être utiles pour des applications pédagogiques est l'enrichissement automatique (ou sous forme de recommandation) des transcriptions avec des ressources externes. C'est l'un des objectifs des applications pédagogiques qui vise à offrir aux étudiants des liens externes utiles qui peuvent servir pour réviser ou avoir plus d'explications détaillées concernant les concepts du cours. Dans ce cas, il est important d'évaluer l'impact de la transcription sur une tâche de récupération de documents. Notre évaluation consiste à comparer les résultats de requêtes de recherche pour chaque segment de « Granularité 1 ». Les requêtes sont construites en utilisant l'approche TF-IDF sur les transcriptions de chacun de ces segments. Ces requêtes sont soumises à un moteur de recherche (ici, Google). Notre but est de déterminer la pertinence des documents récupérés. Nous avons défini comme documents pertinents (référence) les documents extraits de requêtes basées sur les transcriptions

manuelles. Sur la base de cette référence, une comparaison avec les documents récupérés à partir de requêtes construites sur des transcriptions résultant d'une reconnaissance automatique sans et avec adaptation des modèles de langage a été effectuée, en calculant un taux de couverture.

### 4.2 Méthodologie d'évaluation : tâche d'indexation

Dans cette seconde tâche, notre objectif est d'évaluer l'indexabilité des transcriptions. En d'autres termes, nous souhaitons déterminer si la qualité des transcriptions joue un rôle dans l'indexation et la récupération des transcriptions, qui sont utiles pour naviguer dans la vidéo du cours et pour atteindre rapidement ce que cherche l'apprenant. Les segments de type « Granularité 1 » ont été indexés en utilisant le moteur de recherche lemur [1]. Trois ensembles de segments ont été considérés : ceux des transcriptions manuelles, ceux des transcriptions automatiques sans adaptation et ceux des transcriptions automatiques avec adaptation. Des requêtes vont être présentés au moteur de recherche (lemur) pour récupérer les segments pertinents à partir de chaque ensemble distinct de segments. Les requêtes utilisées sont les 40 mots les plus fréquents dans la transcription manuelle, les mots des titres des diapositives et les mots clés annotés à partir de la transcription manuelle. Chaque requête renvoie une liste ordonnée de segments. Pour estimer la qualité de l'indexabilité, nous avons utilisé le coefficient de Spearman qui mesure la corrélation de rang entre les segments récupérés à partir d'une recherche dans l'ensemble de segments de la transcription manuelle et les segments récupérés à partir d'une recherche dans l'ensemble de segments de la transcription automatique (sans et avec adaptation).

## 5 Résultats et discussions

### 5.1 Performances en WER

Le tableau 2 présente les résultats d'adaptation du modèle de langage en utilisant la métrique WER. Nous remarquons une réduction absolue de 3,04% en WER avec un système adapté au domaine.

| | **SRAP sans adaptation** | **SRAP avec adaptation** |
|---|---|---|
| Taux d'erreurs sur les mots (WER) | 19,46 | 16,42 |

TABLE 2 – Résultats d'adaptation du modèle de langage en WER

### 5.2 Performances pour la tâche de recherche d'information

La figure 1 présente le taux de couverture des documents récupérés à partir de requêtes construites sur des transcriptions automatiques, avec (lignes continues) ou sans (pointillés) adaptation des modèles de langage, par rapport aux documents récupérés à partir de requêtes construites sur des transcriptions manuelles. Le taux de couverture est calculé en fonction du nombre de documents visés (de 1 à 20). Nous avons également expérimenté différents types de requêtes composées de 1 à 5 mots (k dans la figure 1) extraits par TF-IDF. Les résultats montrent que la transcription avec adaptation surpasse la

1. https ://www.lemurproject.org

transcription sans adaptation en termes de récupération des ressources pertinentes, pour toutes les tailles de requêtes.

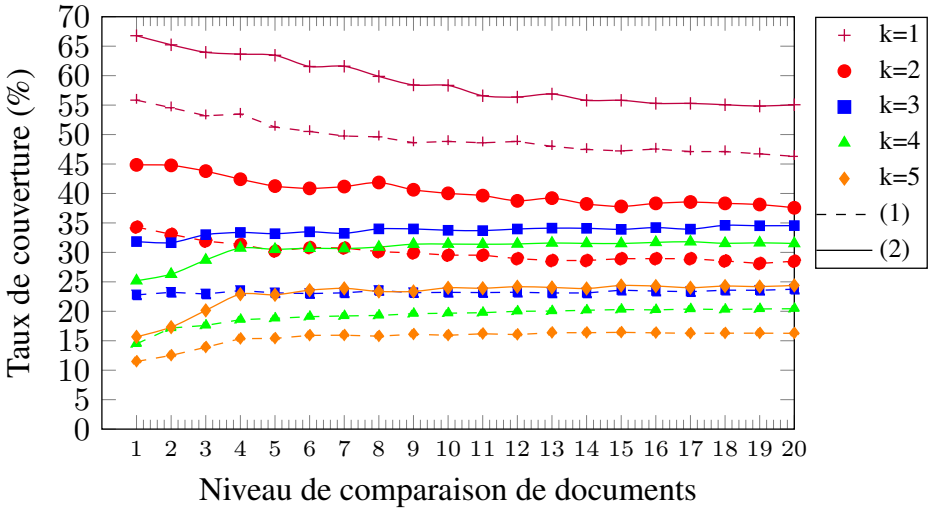



FIGURE 1 – Tâche de recherche d'information : comparaison du taux de couverture entre les requêtes construites à partir de segments de transcriptions manuelles et de transcriptions automatiques (1) sans et (2) avec adaptation des modèles de langage.

## 5.3 Performances pour la tâche d'indexation

Le tableau 3 présente les scores de corrélation moyens pour l'ensemble du corpus en utilisant trois ensembles de jeux de requête : les 40 mots les plus fréquents de la transcription, les mots des titres des diapositives, les mots clés de la transcription. Le coefficient de Spearman varie entre -1 et +1. Une valeur proche de +1 indique une forte corrélation entre les deux listes de documents renvoyés par la recherche alors qu'une valeur proche de 0 indique une faible corrélation (-1 indique une forte corrélation mais dans un sens opposé). Le tableau 3 présente le score moyen de corrélation pour l'ensemble du corpus. Ici également, les résultats indiquent une meilleure indexabilité obtenue après adaptation du modèle de langage du SRAP, pour tous les jeux de requêtes.

| **Jeux de requêtes** | **SRAP sans adaptation** | **SRAP avec adaptation** |
|---|---|---|
| Les 40 mots les plus fréquents de la transcription manuelle | 0,367 | 0,498 |
| Les mots des titres des diapositives | 0,458 | 0,588 |
| Les mots clés annotés à partir de la transcription manuelle | 0,288 | 0,516 |

TABLE 3 – Évaluation de l'indexabilité des transcriptions : comparaison des résultats d'extraction avec le coefficient de corrélation de rang de Spearman, en utilisant différents jeux de requêtes

## 5.4 Discussions

Comme nous l'avons vu dans notre cadre expérimental, l'adaptation automatique de modèles de langage pour la reconnaissance de la parole permet de réaliser environ trois erreurs de moins pour

cent mots transcrits (WER passant de 19,46% à 16,42%), ce qui correspond à une réduction du WER d'environ 15,6%. Ces valeurs, bien qu'intéressantes, ne mettent pas en avant certains phénomènes très intéressants liés aux tâches finales pour lesquelles les transcriptions automatiques sont générées.

En termes de recherche d'information, par exemple, nous constatons une augmentation du taux de couverture des documents retrouvés (par rapport aux documents qui auraient été trouvés à partir de requêtes extraites de transcriptions manuelles) qui peut dépasser 28,5% (k=1, niveau 1, taux de couverture passant de 56% à 67%). Enfin, en termes d'indexabilité, nous montrons que, dans cette étude, le taux de corrélation de Spearman (par rapport à l'indexation obtenue par des transcriptions manuelles) peut augmenter de plus de 79% (de 0,288 à 0,516) pour les termes les plus importants du document grâce à l'adaptation des modèles de langage.

Les résultats présentés montrent que le WER ne permet pas de bien mesurer les performances pour les tâches de recherche d'information et d'indexation considérées, puisqu'il masque les gains réels apportés par l'adaptation des modèles de langage sur les tâches visées : cela prouve la nécessité d'utiliser de nouvelles mesures, telles que celles présentées, pour évaluer l'apport réel de l'adaptation des modèles de langage.

# 6 Conclusion

Le taux d'erreurs sur les mots WER n'est pas toujours la meilleure mesure à utiliser pour évaluer les systèmes de reconnaissance de la parole. Une meilleure compréhension de l'impact des erreurs de transcription automatique implique nécessairement le développement de meilleures mesures (ou métriques) pour l'évaluation, afin de prendre en compte le contexte applicatif dans lequel les SRAP sont utilisés. Dans cet article, nous avons proposé l'utilisation de deux mesures d'évaluation extrinsèque pour évaluer la capacité de créer des requêtes pertinentes pour l'extraction d'information et l'indexabilité de transcriptions automatiques. Nous avons appliqué ces méthodes d'évaluation pour mesurer l'impact de l'adaptation des ML dans le contexte de cours magistraux. Les résultats obtenus montrent que le taux d'erreur sur les mots est une métrique insuffisante qui masque les apports effectifs de l'adaptation des modèles de langage.

# Remerciements

Nous remercions l'agence ANR pour son financement à travers le projet PASTEL sous le numéro de contrat ANR-16-CE33-0007.

# Références

BELL P., YAMAMOTO H., SWIETOJANSKI P., WU Y., MCINNES F., HORI C. & RENALS S. (2013). A lecture transcription system combining neural network acoustic and language models. In *Proc. of the Annual Conference of the International Speech Communication Association (Interspeech'13)*, p. 3087–3091.

CERVA P., SILOVSKY J., ZDANSKY J., NOUZA J. & MALEK J. (2012). Real-time lecture transcription using asr for czech hearing impaired or deaf students. In *Proc. of the Annual Conference of the International Speech Communication Association (Interspeech'12)*.

FAVRE B., CHEUNG K., KAZEMIAN S., LEE A., LIU Y., MUNTEANU C., NENKOVA A., OCHEI D., PENN G., TRATZ S. *et al.* (2013). Automatic human utility evaluation of ASR systems : Does WER really predict performance ? In *Proc. of the Annual Conference of the International Speech Communication Association (Interspeech'13)*, p. 3463–3467.

HÜRST W., KREUZER T. & WIESENHÜTTER M. (2002). A qualitative study towards using large vocabulary automatic speech recognition to index recorded presentations for search and access over the web. In *Proc. of the International Conference on WWW/Internet (ICWI'02)*, p. 135–143.

JANNET M. A. B., GALIBERT O., ADDA-DECKER M. & ROSSET S. (2015). How to evaluate ASR output for named entity recognition ? In *Proc. of the Annual Conference of the International Speech Communication Association (Interspeech'15*.

KAWAHARA T., NEMOTO Y. & AKITA Y. (2008). Automatic lecture transcription by exploiting presentation slide information for language model adaptation. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'08)*, p. 4929–4932.

LECORVÉ G., GRAVIER G. & SÉBILLOT P. (2008). An unsupervised web-based topic language model adaptation method. In *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'08)*, p. 5081–5084.

LUZZATI D., GROUIN C., VASILESCU I., ADDA-DECKER M., BILINSKI E., CAMELIN N., KAHN J., LAILLER C., LAMEL L. & ROSSET S. (2014). Human annotation of ASR error regions : Is" gravity" a sharable concept for human annotators ? In *Proc. of the International Conference on Language Resources and Evaluation (LREC'14)*, p. 3050–3056.

MARTÍNEZ-VILLARONGA A., MIGUEL A., ANDRÉS-FERRER J. & JUAN A. (2013). Language model adaptation for video lectures transcription. In *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'13)*, p. 8450–8454.

MDHAFFAR S., LAURENT A. & ESTÈVE Y. (2018). Le corpus PASTEL pour le traitement automatique de cours magistraux. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN'18)*.

PALLETT D. S. (2003). A look at NIST's benchmark ASR tests : Past, Present, and Future. In *Proc. of the Workshop on Automatic Speech Recognition and Understanding (ASRU'03)*, p. 483–488.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLICEK P., QIAN Y., SCHWARZ P. *et al.* (2011). The Kaldi speech recognition toolkit. In *Proc. of the Workshop on Automatic Speech Recognition and Understanding (ASRU'11)*.

POVEY D., PEDDINTI V., GALVEZ D., GHAHREMANI P., MANOHAR V., NA X., WANG Y. & KHUDANPUR S. (2016). Purely sequence-trained neural networks for ASR based on lattice-free mmi. In *Proc. of the Annual Conference of the International Speech Communication Association (Interspeech'16)*, p. 2751–2755.

ROUSSEAU A., BOULIANNE G., DELÉGLISE P., ESTÈVE Y., GUPTA V. & MEIGNIER S. (2014). LIUM and CRIM ASR system combination for the REPERE evaluation campaign. In *Proc. of the International Conference on Text, Speech, and Dialogue (TSD'14)*, p. 441–448.

YAMAZAKI H., IWANO K., SHINODA K., FURUI S. & YOKOTA H. (2007). Dynamic language model adaptation using presentation slides for lecture speech recognition. In *Proc. of the Annual Conference of the International Speech Communication Association (ICSLP'07)*.

# Apprentissage faiblement supervisé de la structure discursive

Sonia Badene [1,2] Kate Thompson [1] Jean-Pierre Lorré [2] Nicholas Asher [1]
(1) IRIT, Université de Toulouse, (2) LINAGORA
soniabadene@gmail.com, kate.thompson@irit.fr, jplorre@linagora.com,
nicholas.asher@irit.fr

RÉSUMÉ

L'avènement des techniques d'apprentissage automatique profond a fait naître un besoin énorme de données d'entraînement. De telles données d'entraînement sont extrêmement coûteuses à créer, surtout lorsqu'une expertise dans le domaine est requise. L'une de ces tâches est l'apprentissage de la structure sémantique du discours, tâche très complexe avec des structures récursives avec des données éparses, mais qui est essentielle pour extraire des informations sémantiques profondes du texte. Nous décrivons nos expérimentations sur l'attachement des unités discursives pour former une structure, en utilisant le paradigme du *data programming* dans lequel peu ou pas d'annotations sont utilisées pour construire un ensemble de données d'entraînement "bruité". Le corpus de dialogues utilisé illustre des contraintes à la fois linguistiques et non-linguistiques intéressantes qui doivent être apprises. Nous nous concentrons sur la structure des règles utilisées pour construire un modèle génératif et montrons la compétitivité de notre approche par rapport à l'apprentissage supervisé classique.

ABSTRACT

**Learning discourse structure using weak supervision**

The advent of Deep Learning techniques has created a critical need for more labeled training data. Such training data are extremely expensive to create, especially when domain expertise is required. One such task is the learning of semantic structure and discourse structure, which are typically very complex involving recursion but which are essential for extracting deep semantic information from text. In this article, we will show our experiments with the *data programming* paradigm, in which few to no annotations are used to build the data set for the attachment problem in discourse, the first step in forming the complete structure of discourse. The corpus of situated dialogues we use exhibits interesting structural constraints on both linguistic and non-linguistic components that need to be learned. We focus on the rules structure used to build the generative model and show the competitiveness of our approach compared to traditional supervised learning.

MOTS-CLÉS : Structure du discours, Supervision distante, Attachement, *Data Programming* .

KEYWORDS: Discourse Structure, Weak Supervision, Attachment, Data Programming .

# 1 Introduction

L'arrivée des nouvelles méthodes d'apprentissage profond a beaucoup simplifié l'étape d'extraction de caractéristiques de données. Cependant, pour que ces techniques basées sur des algorithmes d'apprentissage supervisé puissent apprendre ces attributs, il faut avoir beaucoup de données étiquetées, des données d'apprentissage sur lesquelles ils peuvent être formés. L'étiquetage manuel des données est à la fois coûteux et long, surtout lorsqu'une expertise dans le domaine est requise ou qu'ultérieurement il faut ré-étiqueter d'une nouvelle manière les données.

Dans cet article, nous montrerons comment Snorkel (Ratner *et al.*, 2017), un nouveau paradigme de programmation par les données (*data programming* en anglais) peut construire un ensemble suffisant de données d'entraînement pour résoudre l'attachement, un problème clef dans le processus d'inférence d'une structure pour un discours. La méthode de Snorkel ne perd que 4% d'exactitude par rapport à une méthode classique d'entraînement avec des données manuellement annotées.

Plus important encore, nous montrons que dans le cadre de Snorkel on peut construire un "modèle génératif" qui est en effet *plus performant* que notre modèle classique. Les probabilités du modèle génératif peuvent aussi servir d'entrée à un modèle discriminatif. Snorkel apporte un cadre pour fournir des informations symboliques de manière efficace à un processus connexionniste ou statistique qui doit généraliser et lisser les résultats fournis par cette partie symbolique, exemplifiant ainsi une IA "hybride" employant des représentations symboliques et des méthodes connexionnistes.

# 2 État de l'art

Il existe plusieurs théories de la structure du discours pour les textes : RST (Rhetorical Structure Theory) (Mann & Thompson, 1987), LDM (Linguistic Discourse Model) (Polanyi *et al.*, 2004), PDTB (The Penn Discourse Treebank) (Prasad *et al.*, 2007) et SDRT (Segmented Discourse Representation Theory) (Asher & Lascarides, 2003). Bien que le travail d'analyse du discours soit largement concentré sur la théorie de la structure rhétorique (RST), comme démontré par Morey *et al.* (2018), les structures discursives ont intérêt à être traduites dans des arbres de dépendance (Muller *et al.*, 2012; Afantenos *et al.*, 2015). De plus, nous nous intéressons à l'étude de dialogues multi-locuteurs pour lesquels le seul corpus annoté est STAC [1] (Asher *et al.*, 2016). Dans ce corpus on trouve une portion significative de structures qui sont naturellement interprétées de façon non arborescente, ce qui exclurait un traitement en termes de DLTAG, LDM ou RST. Nous partons sur la base de cette discussion et travaillons avec des structures de la SDRT simplifiées et nous nous baserons sur les initiatives de Perret *et al.* (2016) où les structures discursives attendues sont des graphes. [2]

Dans cet article, nous proposons la création des données d'entraînement pour la tâche de prédiction d'attachement dans le corpus de dialogues STAC (Asher *et al.*, 2019). Pour cela, nous utilisons la *programmation par les données*, un paradigme pour la création et la modélisation des données d'entraînement. La programmation par les données fournit un cadre simple et unificateur pour une faible supervision, dans lequel les étiquettes d'entraînement sont bruitées et peuvent provenir de sources multiples et potentiellement contradictoires. Dans ce cadre, on peut coder cette faible supervision sous la forme de fonctions d'étiquetage, qui fournissent chacune une étiquette pour un sous-ensemble de données. De nombreuses approches de supervision faibles différentes peuvent être exprimées sous forme de fonctions d'étiquetage, telles que les stratégies qui utilisent des bases de connaissances existantes, comme dans la supervision distante (Mintz *et al.*, 2009).

# 3 Expérimentations

## 3.1 Corpus de dialogues annotés

Pour notre expérimentation nous avons utilisé STAC, un corpus de discussions spontanées autour du jeu *Colons de Catan* de négociations multi-locuteurs annotées pour la structure discursive dans le style

1. Lien vers le corpus STAC : https ://www.irit.fr/STAC/index.html

2. Les structures de la SDRT ont une complexité difficile à gérer–les structures complexes ou CDUs (Complexe Discourse Units) (Venant *et al.*, 2013). Comme d'autres travaux sur les structures discursives en SDRT (Muller *et al.*, 2012; Perret *et al.*, 2016), nous simplifions cette structure des CDUs (voir Section 3.2).

de la SDRT. Cette annotation inclut des coups linguistiques mais aussi des actions non-linguistiques comme l'action de terminer son tour de négociation ou de construire une route comme illustrée sur la figure 1. Nous avons choisi ce corpus qui est annoté manuellement sur l'attachement afin d'évaluer notre approche, mais aussi parce que l'analyse des corpus de dialogues est de plus en plus demandée avec l'arrivée des assistants conversationnels, des chatbots ou des corpus de plateforme de discussion en ligne.

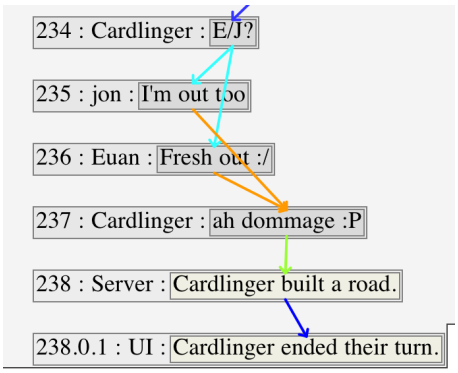



FIGURE 1 – Extrait d'un dialogue sur STAC qui montre des relations comme Sequence (bleu foncé), Result (vert), et les coups linguistiques (énoncés par les joueurs) et non-linguistiques (donnés par le "Serveur" ou "UI")

Le corpus présente des contraintes structurelles intéressantes sur les composantes linguistiques et non linguistiques qui doivent être apprises. Le corpus non-linguistique est très régulier, donc facile à modéliser. Par contre, la structure discursive entre tours linguistiques, ou entre tours non-linguistique et tours linguistiques est riche et difficile à capturer. Dans notre étude, on s'intéresse aux tours linguistiques et aux tours qui marquent une transition entre le linguistique et les coups non-linguistiques.

## 3.2 Cadre expérimental

Avant de commencer nos expériences, nous avons mis en œuvre les prétraitements suivants :

1. La complexité des structures annotées de la SDRT étant difficile à prédire, nous avons suivis les travaux de (Muller *et al.*, 2012; Perret *et al.*, 2016) et avons mis en place un algorithme simple de "flattening" ("aplatissement") afin de remplacer les CDUs par des relations entre paires d'unités élémentaires discursives. Un CDU est un graphe de dépendance avec plusieurs DU/segments comme sommets. Cet ensemble est considéré comme un segment qui peut être reliée à d'autres segments. Pour aplatir les graphes, pour chaque CDU nous avons identifié la "tête" du CDU, qui est le premier segment du CDU, et connectons toutes les relations entrantes et sortantes du CDU à la tête - pour nos 4 types de relations, il y a un total de 33 865 relations dans le corpus (incluant celles des dialogues non linguistiques), et environ 40% de celles-ci ont été ajustées, soit au noeud source, soit au noeud cible (ou aux deux).
2. Nous avons également restreint les dialogues dans le corpus afin d'étudier essentiellement les dialogues linguistiques. Nous avons éliminer tous les dialogues qui n'ont pas de conversations linguistiques - 1463 dialogues ont été supprimés, ce qui nous laisse 1130 dialogues qui contiennent 18 767 segments non linguistiques et 13 734 segments linguistiques, avec 31 251 relations.

3. Nous avons aussi ignoré tous les attachements qui ont une distance supérieure à 10 (c'est-à-dire qui ont plus de 9 unités élémentaires discursives entre la source et la cible de l'attachement). Les distances relationnelles varient de 1 à 160 dans le corpus de développement. 67% des relations ayant la distance 1, et 98% des relations ayant la distance 10 ou moins.
4. Pour cette première tâche de prédiction d'attachement, nous avons travaillé avec 4 types de relations les plus fréquentes : Question-answer-pair (QAP), Sequence (temporelle), Result (relation causale), Continuation (la relation de continuité thématique) - ce qui représente 70% des relations dans le corpus.
5. Afin de réduire le temps d'exécution de chaque règle pendant le développement, nous avons créé des ensembles restreints pour chaque type de relation : des versions plus petites de l'ensemble de développement qui ignorent toutes les paires qui ne pouvaient pas être attachées par le type de relation en question. Nous disposons de sous-ensembles de données pour chaque relation discursive particulière et d'un ensemble plus vaste des données pour les règles des quatre relations discursives que nous avons examinées.

## 3.3 Les différentes étapes du traitement

### 3.3.1 Les candidats et les Fonctions d'étiquetage

Les candidats sont les entrées pour lesquelles les étiquettes seront prédites et ils sont extraits des données en fonction de la nature de la tâche de prédiction. Puisque nous prédisons l'attachement entre deux segments d'un dialogue, nos candidats sont l'ensemble de toutes les combinaisons possibles *source-cible* des segments dans chaque dialogue, limitées par une distance relationnelle maximale de 10. Le corpus STAC nous donne déjà la segmentation en unités discursives élémentaires des dialogues du corpus. Nous avons construit notre propre algorithme pour générer les candidats en entrée pour notre modèle génératif. Pour chaque dialogue, nous avons créé une liste de toutes les paires uniques possibles. Comme certaines relations discursives présentent des liens vers l'arrière (les 4 relations étudiées dans cet article n'ayant pas de relations vers l'arrière, nous avons tout de même généré ces candidats afin d'évaluer l'attachement sur la structure globale), nous avons ajouté des contraintes comme dans (Perret *et al.*, 2016) afin d'éviter l'explosion combinatoire de candidats.

Dans Snorkel, les règles s'appellent des fonctions de labellisation (FL). Les FL sont appliquées à chaque candidat et retourne un "1", "-1" ou "0" qui signifie que les deux segments du candidat sont "liés" , "non-liés", ou "on ne sait pas". Ces fonctions utilisent les informations "locales" des candidats (le texte avec la syntaxe, les connecteurs ...), y compris les identités des interlocuteurs et des destinataires, les actes de dialogue, les types des segments (linguistique ou non-linguistique) et la distance entre les segments, afin de saisir le modèle général sous-tendant. Comme nous l'avons vu plus haut, nous prédisons l'attachement en aillant en tête les 4 types de relations – *Result*, *QAP*, *Continuation* et *Sequence*. De cette façon, les FL utilisent également une information de type. Cela a du sens d'un point de vue à la fois empirique et épistémologique : une décision d'attachement discursive entre deux segments est étroitement liée au type de la relation qui les lie, et donc lorsqu'un annotateur décide que deux éléments discursifs élémentaires sont attachés, il ou elle le fait avec une certaine connaissance du type de relation qui les lie. La figure 2 montre un exemple de nos règles utilisées pour la prédiction d'attachement avec la relation *Result* en tête.

Si nous nous concentrons sur l'information locale en construisant les FL, le besoin de s'appuyer sur l'information considérée globale devient évident du fait que, si nous n'avons pas de moyen de surveiller et noter où nous sommes exactement dans un dialogue et où sont les attachements déjà

```python
def LF_Result_L_L_case1(row):
    l=0
    if (any(x in row.target_text.lower() for x in resultWords)
                or any(x in row.source_text.lower() for x in resultWords)):
            l=1
    return l


def LF_Result_L_L_case2(row):
    l = 0
    if row.source_surface_act in ["Question", "Request", "Assertion"] \
    and (row.target_dialogue_act in ["Offer", "Counteroffer"] \
         or row.source_emitter == row.target_text.partition(' ')[0] or row.target_surface_act == "Request"):
        l=1
    return l
```

FIGURE 2 – *Result* relie une cause à son effet. Voici un exemple de nos règles écrites en python pour la relation *Result* reliant deux unités de discours linguistiques.

prédits, nous risquons de sur-étiqueter les candidats “liés”. Ce qui est très inefficace dans un corpus dont les données sont éparses. Ainsi, nous avons ordonné les candidats pour appliquer les FL aux candidats des segments adjacents en premier, et puis regarder ceux des segments de plus en plus éloignés, tout en maintenant une liste de tous les segments déjà prédits “lié”. Ces mesures prises nous permettent de nous servir des faits contextuels simples, et donc de construire des FL plus sensibles au contexte, ce qui aboutit à une différence de 5 points d’exactitude de nos règles par rapport aux exemples dans le corpus de développement.

### 3.3.2 Le modèle génératif

Une fois que nous appliquons l’ensemble des FL à tous les candidats, nous passons à l’étape "générative". Dans le système Snorkel, le modèle génératif unit les résultats des FL : une matrice des étiquettes donnée par chaque FL (colonnes) sur les candidats (lignes) est alors générée. Bien que l’approche la plus simple serait de prendre le vote majoritaire entre les FL pour chaque candidat, celle-ci serait moins efficace dans les situations où nous n’avons pas beaucoup de votes sur une entrée, ou si toutes les FL s’abstiennent. De plus, elle ne prendrait pas en compte les performances individuelles des FL. Donc pour apporter des améliorations au vote majoritaire, le modèle génératif cherche à maximiser la probabilité marginale des FL de chaque candidat pour apprendre une estimation des précisions des FL et les pondérer selon ces précisions (Bach *et al.*, 2017). Ensuite, le modèle calcule pour chaque candidat la probabilité d’être “1” ou “0” (“lié” ou “non-lié”) dans le contexte de notre tâche de prédiction binaire.

Ce calcul suppose que les FL sont indépendantes. Cependant, les FL fournies sont souvent dépendantes : par exemple, les FL peuvent être de simples variations les unes des autres ou peuvent dépendre d’une source commune de supervision distante. Si nous ne tenons pas compte des dépendances entre les fonctions d’étiquetage, nous pouvons avoir toutes sortes de problèmes. La méthode de sélection automatique des dépendances à modéliser dans Snorkel, sans accès aux données de référence, utilise un estimateur de pseudo-vraisemblance, qui ne nécessite aucun échantillonnage ni approximation pour calculer le gradient objectif et ceci est plus rapide que l’estimation du maximum de vraisemblance. Cela évite d’indiquer les dépendances à la main, tâche difficile et sujet aux erreurs.

### 3.3.3 Un modèle discriminatif de référence

Alors que le modèle génératif est essentiellement une combinaison pondérée des FL fournies par l'utilisateur - qui ont tendance à être précises mais à faible couverture-, le modèle discriminatif peut conserver cette précision tout en apprenant à généraliser au-delà des fonctions d'étiquetage, augmentant ainsi la couverture et la robustesse sur les données non encore visualisées. Le rappel est alors plus élevé dans la plupart des cas, même si parfois on observe une petite baisse de précision.

Nous avons utilisé le modèle de classification séquentielle de BERT (Devlin *et al.*, 2018) (code source sur le lien ci-dessous [3]) avec 10 époques pour l'entraînement et tous les paramètres par défaut. BERT, Bidirectional Encoder Representations from Transformers, est un "encoder" de texte entraîné à l'aide de modèles de langage où le système doit deviner un mot manquant ou un élément de mot qui est supprimé au hasard du texte. Conçue à l'origine pour les tâches de traduction automatique, BERT utilise l'auto-attention bidirectionnelle pour produire les encodages et produit des résultats qui dépassent l'état de l'art sur de nombreuses tâches de classification textuelle. Alors qu'en principe, nous aurions pu utiliser n'importe quel modèle discriminatif, comme le suggère la littérature de Snorkel, BERT nous a donné de loin les meilleurs résultats sur la prédiction de l'attachement. C'est pourquoi nous avons également utilisé BERT comme modèle pour l'apprentissage supervisé de l'attachement afin de comparer ses résultats avec ceux de la méthode de supervision faible.

# 4 Résultats et analyse

| | VP | VN | FP | FN | Exactitude |
|---|---|---|---|---|---|
| QAP LL | 294 | 1798 | 112 | 138 | 0.89 |
| QAP NLNL | 84 | 187 | 0 | 0 | 1 |
| RES NLNL | 739 | 2929 | 13 | 55 | 0.98 |
| RES LNL | 13 | 2158 | 93 | 97 | 0.91 |
| RES LL | 25 | 316 | 19 | 37 | 0.85 |
| RES NLL | 2 | 139 | 0 | 2 | 0.98 |
| Cont LL | 16 | 9818 | 110 | 106 | 0.97 |
| Cont NLNL | 613 | 3254 | 0 | 1 | 0.99 |
| SEQ NLL | 90 | 658 | 2 | 14 | 0.97 |
| SEQ NLNL | 236 | 1220 | 10 | 76 | 0.94 |

TABLE 1 – Nombre de vrais positifs (VP), de vrais négatifs(VN), de faux positifs (FP) et de faux négatifs (FN) pour chacune de nos fonctions d'étiquetage lorsqu'elles sont appliqués aux candidats associés aux types des relations discursives utilisées.

Nous avons d'abord évalué les FL pour chaque type de relation discursive individuellement sur les sous-corpus de développement, en fournissant une mesure de leur couverture et de leur précision (Tableau 1). Ensuite, nous avons évalué le modèle génératif sur la combinaison des quatre types de FL. Le tableau 2 présente les résultats à la fin de chaque étape de notre système de supervision faible (Modèle Génératif). Pour comparer les deux approches, le modèle discriminatif a été entraîné sur les marginaux fournis par notre modèle génératif, mais aussi sur les annotations manuelles du corpus Stac. L'évaluation du modèle discriminatif s'est faite sur l'ensemble test du corpus.

Les résultats du modèle génératif sur l'attachement sont près de 20 points plus élevés en F1 mesure

3. Lien vers le code source du modèle de classification séquentielle de BERT : https ://github.com/huggingface/pytorch-pretrained-BERT/blob/master/examples/run_classifier.py

| | **Modèle Génératif** | | | **Modèle Discriminatif sur Test** | |
|---|---|---|---|---|---|
| | Dev | Train | Test | avec Marginales | avec annotations Manuelles |
| Précision | 0.67 | 0.70 | 0.68 | 0.45 | 0.61 |
| Rappel | 0.84 | 0.85 | 0.84 | 0.54 | 0.53 |
| F1 mesure | 0.75 | 0.77 | 0.75 | 0.49 | 0.57 |
| Exactitude | 0.92 | 0.93 | 0.92 | 0.84 | 0.88 |

TABLE 2 – Évaluations de prédiction de l'attachement avec la combinaison de toutes les règles des quatre types modélisées dans cet article, avec les approches faiblement supervisées et supervisées.

par rapport au modèle discriminatif entraîné sur les annotations manuelles. Cela montre la puissance de l'approche fondée sur des règles et la supervision faible, même lorsqu'on la compare à un système d'apprentissage profond à l'état de l'art. Un autre point intéressant est que le modèle discriminatif a des résultats acceptables avec les données marginales par rapport à sa performance en utilisant les annotations manuelles ; son exactitude n'est inférieure que de 4 points et son score F1 est inférieur de 8 points mais toujours comparable aux résultats de la littérature, montrant que le modèle génératif transmet bien des informations au modèle discriminatif. Plutôt que de traiter naïvement ces étiquettes bruyantes comme une vérité de base, notre modèle discriminatif sensible au bruit donne une légère amélioration dans le rappel avec une diminution de la précision par rapport à l'approche supervisée. En ce qui concerne les FL individuelles isolées, nous constatons qu'à part *QAP*, nos règles pour chaque type de relation ont une exactitude, une précision et un rappel comparables à ceux des modèles supervisés. L'une des raisons de notre précision plus faible pour *QAP* peut être attribuée aux conséquences de la procédure d'aplatissement réalisée en pré-traitement ; dans certains cas, l'algorithme d'aplatissement rattache la relation *QAP* à la tête d'un CDU qui en fait n'était pas le segment du CDU qui a marqué la question. Ce qui est intéressant, c'est la synergie entre les règles, de sorte que lorsqu'elles interagissent toutes sur les données de test, elles réussissent très bien sur le modèle génératif.

# 5 Conclusion et perspectives

Ayant choisi un modèle discriminatif unique pour toutes nos expérimentations, nous avons pu comparer notre approche hybride utilisant Snorkel avec celle du modèle classique sur une tâche difficile, celle de l'attachement discursive. Notre approche permet de modéliser plus finement le discours et d'être généralisée à d'autres corpus. Contrairement à un algorithme supervisé, nos résultats sur le modèle génératif sont supérieurs de près de 30 points sans couvrir tous les types de règles. Nous générons ainsi beaucoup de données annotées en très peu de temps. Comme perspectives nous envisageons d'enrichir notre modèle Snorkel d'abord en couvrants tous les types de relations et en implémentant des règles qui prendront en compte les contraintes de structuration globale, et non seulement au niveau des paires d'éléments discursives élémentaires comme réalisé jusqu'à présent.

# Remerciements

Ce travail a été réalisé dans le cadre du projet de recherche PIA Grands Défis du Numérique LinTO-Assistant vocal open-source respectueux des données personnelles pour l'entreprise- soutenu par Bpifrance N°P169201.

# Références

AFANTENOS S., KOW E., ASHER N. & PERRET J. (2015). : Association for Computational Linguistics (ACL).

ASHER N., HUNTER J., MOREY M., BENAMARA F. & AFANTENOS S. D. (2016). Discourse structure and dialogue acts in multiparty dialogue : the stac corpus. In *LREC*.

ASHER N., HUNTER J. & THOMPSON C. (2019). Comparing discourse structures between purely linguistic and situated messages in an annotated corpus. submitted.

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

BACH S. H., HE B., RATNER A. & RÉ C. (2017). Learning the structure of generative models without labeled data. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, p. 273–282 : JMLR. org.

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

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

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

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

MANN W. C. & THOMPSON S. A. (1987). Rhetorical structure theory : Description and construction of text structures. In *Natural language generation*, p. 85–95. Springer.

MINTZ M., BILLS S., SNOW R. & JURAFSKY D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP : Volume 2-Volume 2*, p. 1003–1011 : Association for Computational Linguistics.

MOREY M., MULLER P. & ASHER N. (2018). A dependency perspective on rst discourse parsing and evaluation. *Computational Linguistics*, p. 198–235.

MULLER P., AFANTENOS S., DENIS P. & ASHER N. (2012). Constrained decoding for text-level discourse parsing. *Proceedings of COLING 2012*, p. 1883–1900.

PERRET J., AFANTENOS S., ASHER N. & MOREY M. (2016). Integer linear programming for discourse parsing. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 99–109.

POLANYI L., CULY C., VAN DEN BERG M., THIONE G. L. & AHN D. (2004). A rule based approach to discourse parsing. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*.

PRASAD R., MILTSAKAKI E., DINESH N., LEE A., JOSHI A., ROBALDO L. & WEBBER B. (2007). The penn discourse treebank 2.0. annotation manual. the pdtb research group.

RATNER A., BACH S. H., EHRENBERG H., FRIES J., WU S. & RÉ C. (2017). Snorkel : Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, **11**(3), 269–282.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

VENANT A., ASHER N., MULLER P., DENIS P. & AFANTENOS S. (2013). Expressivity and comparison of models of discourse structure. In *Proceedings of the SIGDIAL 2013 Conference*, p. 2–11.

# CALOR-QUEST : un corpus d'entraînement et d'évaluation pour la compréhension automatique de textes

Frédéric Béchet[1] Cindy Aloui[1] Delphine Charlet[2] Géraldine Damnati[2]
Johannes Heinecke[2] Alexis Nasr[1] Frédéric Herlédan[2]
(1) Aix-Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
(2) Orange Labs, Lannion
(1) {first.last}@lis-lab.fr
(2) {first.last}@orange.com

## RÉSUMÉ

La compréhension automatique de texte est une tâche faisant partie de la famille des systèmes de *Question/Réponse* où les questions ne sont pas à portée générale mais sont liées à un document particulier. Récemment de très grand corpus (SQuAD, MS MARCO) contenant des triplets (document, question, réponse) ont été mis à la disposition de la communauté scientifique afin de développer des méthodes supervisées à base de réseaux de neurones profonds en obtenant des résultats prometteurs. Ces méthodes sont cependant très gourmandes en données d'apprentissage, données qui n'existent pour le moment que pour la langue anglaise. Le but de cette étude est de permettre le développement de telles ressources pour d'autres langues à moindre coût en proposant une méthode générant de manière semi-automatique des questions à partir d'une analyse sémantique d'un grand corpus. La collecte de questions *naturelle* est réduite à un ensemble de validation/test. L'application de cette méthode sur le corpus CALOR-Frame a permis de développer la ressource CALOR-QUEST présentée dans cet article.

## ABSTRACT

Machine reading comprehension is a task related to Question-Answering where questions are not generic in scope but are related to a particular document. Recently very large corpora (SQuAD, MS MARCO) containing triplets (document, question, answer) were made available to the scientific community to develop supervised methods based on deep neural networks with promising results. These methods need very large training corpus to be efficient, however such kind of data only exists for English at the moment. The aim of this study is the development of such resources for other languages by proposing to generate in a semi-automatic way questions from the semantic Frame analysis of large corpora. The collect of *natural questions* is reduced to a validation/test set. We applied this method on the French CALOR-Frame corpus to develop the CALOR-QUEST resource presented in this paper.

MOTS-CLÉS : Compréhension automatique de texte, Question Réponse, Analyse en cadre sémantique, Génération de questions.

KEYWORDS: Machine reading comprehension, Question Answering, Semantic Frame analysis, Question generation.

# 1 Introduction

La compréhension automatique de texte (*Machine Reading Comprehension*) consiste pour une machine à répondre à des questions portant sur un document écrit, chaque réponse se présentant sous la forme d'un passage du document. Il s'agit de la version automatique de la tâche, de *compréhension écrite*, qui permet de vérifier les capacités des élèves à lire un texte et à le comprendre. Du point de vue du Traitement Automatique des Langues il s'agit d'un cas particulier de la tâche de *Question/Réponse* où les questions ne sont pas de portée générale mais sont, au contraire, liées à un document particulier.

Cette tâche a reçu récemment une attention particulière avec la mise à disposition de deux très grands corpus de textes enrichis de questions et réponses : le corpus **SQuAD** (Rajpurkar *et al.*, 2016) et le corpus **MS MARCO** (Nguyen *et al.*, 2016) contenant chacun plus de 100k triplets *(corpus, question, réponse)* où chaque question a été produite par un humain, soit via crowd-sourcing, soit via de vraies requêtes de moteurs de recherche. Ces corpus ont permis le développement de nombreuses approches basées sur de l'apprentissage supervisé, principalement avec des réseaux de neurones profonds, tel que (Wang & Jiang, 2016) ou (Seo *et al.*, 2016), avec une amélioration significative des résultats par rapport aux méthodes fondées sur des analyses linguistiques ou sur des méthodes d'appariement entre questions et texte contenant les réponses (Hermann *et al.*, 2015).

Ces ressources sont disponibles en langue anglaise mais pour d'autres langues, telles que le français, il n'existe pas de corpus comparable et l'effort nécessaire pour collecter une telle quantité de données est très important, limitant l'emploi de ces méthodes à d'autres langues ou à d'autres cadres applicatifs. Pour répondre à ce problème, l'étude présentée dans ce papier vise à créer, de manière partiellement automatique, un corpus permettant d'entraîner de tels systèmes de manière supervisée sans avoir à collecter manuellement une très grande collection de données. La méthode proposée consiste à utiliser une analyse en cadre sémantique de type *FrameNet* sur de grands corpus de texte, puis à générer automatiquement des questions à partir des analyses obtenues. Ces questions automatiques peuvent servir à entraîner des méthodes de compréhension de textes qui seront évaluées sur des questions *naturelles* posées par des évaluateurs humains sur les mêmes documents.

# 2 Travaux similaires

Plusieurs corpus pour la tâche de Compréhension de Texte par la Machine (Machine Reading Comprehension) sont disponibles sous la forme de paires de question/extrait de texte pouvant répondre à la question. Si le corpus **SQuAD** a généré de nombreuses contributions, le corpus **MS-MARCO** produit à partir de requêtes sur le moteur de recherche Bing en est également un exemple récent. On peut se référer à l'article décrivant ce dernier (Nguyen *et al.*, 2016) pour une revue détaillée d'autres corpus en langue anglaise issus de différentes sources comme **NewsQA** (Trischler *et al.*, 2016), **SearchQA** (Dunn *et al.*, 2017) proposant des questions du jeu Jeopardy appariées à des extraits de textes issus de requêtes Google, **NarrativeQA** (Kočiský *et al.*, 2018) construit à partir de résumés de films et de livres. On y trouve également différents types de questions comme des questions à choix multiples dans le corpus **ARC** (Clark *et al.*, 2018) ou des questions insérées sous forme de textes à trous comme dans le corpus **ReCoRD** (Zhang *et al.*, 2018) qui vise à tester l'influence de la connaissance du sens commun. On y trouve également quelques références à des corpus en langue chinoise mais à notre connaissance il n'existe pas de telles ressources en langue française.

Nous cherchons dans cette étude à construire un tel corpus pour le français sans pour autant devoir mettre en oeuvre un processus d'annotation trop complexe. Nous proposons une méthode semi-

automatique partant d'un texte et d'une représentation sémantique de ce texte pour produire les questions associées. La génération de questions à partir d'un texte est une problématique ancienne ayant fait l'objet de nombreux travaux, notamment à l'occasion de campagnes d'évaluation (Boyer & Piwek, 2010). Deux grandes familles de méthodes ont été explorées, que ce soit grâce à des patrons construits à partir de l'analyse syntaxique d'une phrase ou à partir d'une analyse sémantique. Les progrès récents dans ces deux disciplines ont permis de nouvelles avancées en génération de question (Mazidi & Nielsen, 2014). Récemment (Pillai *et al.*, 2018) et (Flor & Riordan, 2018) par exemple proposent de générer des questions factuelles à partir d'une analyse en rôles sémantiques de type PropBank. En revanche ces travaux se situent bien souvent dans un contexte applicatif différent du nôtre, à savoir la production de question pour l'apprentissage d'une langue ou la génération de quizz dans un processus pédagogique. Dans de tels contextes, la lisibilité et la grammaticalité des questions obtenues est primordiale et les questions sont évaluées par des tests subjectifs ou des métriques de type *Bleu* ou *Meteor*. Au delà des approches par règles, des travaux récents envisagent la génération de question comme une tâche d'apprentissage à part entière où la question est générée directement grâce à un réseau de neurones à partir du texte et d'un conditionnement par la réponse (Dong *et al.*, 2018) (Yuan *et al.*, 2017) voire même en envisageant conjointement les tâches de génération de question et de réponse (Wang *et al.*, 2017). Nous proposons une approche de construction de questions à l'aide de patrons reposant sur une analyse sémantique de type FrameNet, nous permettant d'obtenir un corpus utilisable pour étudier la compréhension automatique de texte sur le français.

# 3 Génération semi-supervisée d'un corpus de questions

Le corpus CALOR-Frame est composé de 4 sous-corpus issus de 3 sources encyclopédiques : Wikipédia (WP), Vikidia (V) et ClioTexte (CT). 3 thématiques sont représentées : la première guerre mondiale (1GM), l'archéologie (arch) et l'antiquité (antiq). La variété des sources permet de couvrir différents registres de langage allant du document historique pour ClioTexte (discours, déclarations) aux articles pour enfants dans Vikidia. Ce corpus a été annoté manuellement en cadre sémantique selon le schéma d'annotation *Berkeley FrameNet* (Baker *et al.*, 1998) avec un ensemble de 54 *Frames* décrit dans (Béchet *et al.*, 2017). Les cadres sémantiques ou *Frames* décrivent des situations prototypiques (*décider, perdre, attaquer, vaincre, etc.*). L'annotation consiste d'abord à identifier le mot déclencheur de la Frame, appelé *Lexical Unit (LU)*, puis les actants et circonstants appelés *Frame Elements (FE)*. Le nombre de déclencheurs différents représentés dans le corpus correspond à 145 lemmes (70 noms et 75 verbes). Une même séquence de mots peut correspondre à plusieurs *FEs* différents si une phrase comporte plusieurs *Frames*. Un exemple est donné figure 1 pour une phrase annotée avec les deux *Frames*, `Losing` déclenchée par le mot `perdu` et `Attack` déclenchée par `attaques`.

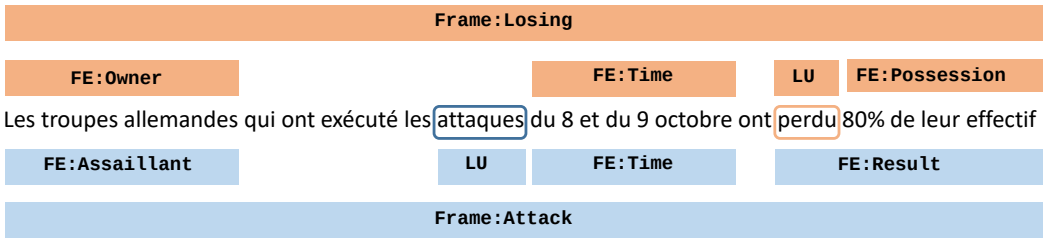



FIGURE 1 – Exemple de phrase annotée avec des cadres sémantiques provenant de FrameNet

Ce type d'annotations est particulièrement intéressant pour l'obtention d'un corpus de *ques-*

*tions/réponses* pour le développement de méthodes de compréhension automatique de texte. En effet, en sélectionnant une *Frame* $F$ et un *Frame Element* $E$ dans une phrase, on peut produire automatiquement des questions dont la réponse est $E$. La Frame ainsi que les autres *FE* présents dans la phrase vont constituer le contexte $C$ de ces questions que l'on va noter sous forme de triplets : $(F, E, C)$. En faisant varier $F$, $E$ et $C$ pour une même phrase, on peut obtenir un ensemble de questions pour lesquelles on dispose des réponses avec leur classe sémantique (le type du *FE*).

Par exemple, sur la phrase de l'exemple 1, nous aurons 18 combinaisons possibles $(F, E, C)$ : `(Losing, Owner, {Time, Possession})`, `(Losing, Owner, {Possession})`, `(Losing, Owner, {Time})`, `(Losing, Time, {Owner, Possession})` ...

A chaque triplet $(F, E, C)$ peut correspondre de nombreuses questions $Q \in$ *Questions*$(F, E, C)$. Par exemple pour la première combinaison, on pourrait avoir : *Qui a perdu 80% de ses effectifs du 8 au 9 octobre ?*, ou encore *Quelles troupes ont été décimées à 80% dans les attaques du 8 et colnou9 octobre ?*. Ces deux questions ont pour réponse le même segment de texte (`les troupes allemandes`), mais ont des formes très différentes : la première est très proche de la phrase originale et pourrait être obtenue par une simple réorganisation de celle-ci sans modification d'ordre lexical, la deuxième en revanche suppose une réécriture complète avec ajout éventuel de nouveaux termes (*décimées*). Ces deux types de questions ont été produits à partir de l'annotation en Frame du corpus CALOR-Frame : une méthode automatique a permis de générer des questions $Q_A$ à partir des combinaisons $(F, E, C)$, et une collecte manuelle sur un sous-ensemble du corpus a permis d'obtenir des questions *naturelles* notées $Q_N$ dans un environnement sémantique contrôlé. Ces deux méthodes sont présentées dans les paragraphes suivants. Dans une première approche, nous nous appuyons sur l'annotation en Frames manuelle du corpus CALOR-Frame pour valider les principes généraux de la méthodologie. Des travaux sont en cours pour valider l'approche à partir de corpora annotés automatiquement par un système d'étiquetage en Frames.

## Production de questions à partir de patrons

La production automatique de questions à partir de cadres sémantiques repose sur des patrons de questions présentés dans la table 1. Un patron est rattaché à un type de Frame $F$. Les mots précédés du symbole \$ dans les patrons de $F$ correspondent aux types de FE de $F$. Les éléments entre crochets sont optionnels et les autres sont obligatoires. De ce fait, un patron peut générer plusieurs questions, en fonction des éléments optionnels choisis.

| | | |
|---|---|---|
| Activity_start | spécifique | Qu'est-ce qui a commencé \$Time [ \$Place ] [ pour \$Purpose ] ?<br>⟶ \$*Activity* . |
| | générique | quoi commencer [ \$Agent ] [ \$Circumstances ] [ \$Co-timed_event ]<br>[ \$Containing_event ] [ \$Event_description ] [ \$Explanation ]<br>[ \$Manner ] [ \$Means ] [ \$Place ] [ \$Purpose ] [ \$Time ] ?<br>⟶ \$*Activity* . |
| Leadership | spécifique | Quand est-ce que \$Leader a dirigé \$Governed [ \$Place ] ?<br>⟶ \$*Time* . |
| | générique | quand diriger [ \$Leader ] [ \$Governed ] [ \$Place ] [ \$Role ]<br>[ \$Duration ] [ \$Activity ] ?<br>⟶ \$*Time* . |

TABLE 1 – Exemples de patrons génériques et spécifiques pour les Frames Activity_start et Leadership

Deux type de patrons ont été développés : des patrons génériques à toutes les Frames, se contentant d'introduire la question par un pronom interrogatif correspondant au type de $E$ puis énumérant toutes

les combinaisons de FE du contexte $C$ ; des patrons spécifiques à chaque Frame, spécifiant les FE obligatoires et facultatifs, ne permettant pas toutes les combinaisons de FE possibles afin de garantir une apparence *naturelle* aux questions générées. Les patrons génériques permettent de générer un très grand nombre de questions couvrant tous les cas possibles, sans se soucier de la justesse syntaxique des phrases générées. Au contraire les patrons spécifiques sont censés générer des phrases plus proches de questions naturelles que pourrait poser un utilisateur.

### Collecter des questions naturelles

La collecte des questions naturelles s'est faite auprès d'annotateurs à qui l'on présentait les éléments $(F, E, C)$. La phrase originale n'était pas affichée pour laisser plus de liberté aux annotateurs dans les choix lexicaux effectués pour rédiger les questions. Les annotateurs avaient aussi toute liberté dans le choix des éléments du contexte $C$ qu'ils allaient inclure dans leurs questions. L'exemple suivant correspond à la configuration `Q(Hiding_objects, Place, {Agent, Hidden_object, Hiding_place})`.

---

**Frame =** `Hiding_objects`
- Contexte
  - **Agent** : `un chef de milice gauloise`
  - **Hidden_object** : `un trésor`
  - **Hiding_place** : `dans sa ferme de Bassing`
- But
  - objet de la question : `Place`
  - réponse : `Moselle`
- Questions collectées :
  - *Dans quelle région un chef de milice gauloise cache-t-il un trésor ?*
  - *Où se trouve la ferme de Bassing, dans laquelle un chef de milice gauloise cache un trésor ?*

---

Les questions naturelles produites avec ce protocole pour la constitution du corpus CALOR-QUEST ont porté sur un sous ensemble du corpus en termes de documents, mais l'ensemble des cadres sémantiques du corpus CALOR-Frame sont représentés.

### Deux ensembles de questions

L'objectif de l'étude n'est pas de valider la qualité intrinsèque des questions générées automatiquement. Si leur construction garantit qu'elles sont valides d'un point de vue sémantique, nous savons qu'elles ne sont pas correctes d'un point de vue syntaxique. Nous ne chercherons donc pas à vérifier si les questions automatiques se retrouvent parmi les questions naturelles. Nous nous intéressons à ce qui relie ces deux types de questions à savoir les occurrences de Frames à partir desquelles elles ont été construites. En effet, une question naturelle et une question automatiques produites à partir d'une même occurrence de Frame et posées sur un même Frame Element ont par construction la même réponse, à savoir ce Frame Element. L'ensemble du corpus avec questions naturelles et questions automatiques sera mis à disposition de la communauté scientifique.

## 4 Evaluation du corpus *CALOR-QUEST*

Pour cette toute première étude sur le corpus CALOR-QUEST, nous évaluons l'hypothèse de travail selon laquelle la réponse à une question naturelle peut être trouvée en effectuant un appariement avec une question générée automatiquement dont on connaît sans ambiguïté la réponse. Pour chaque

question naturelle $Q_N$ d'un document, nous calculons les similarités textuelles avec toutes les questions automatiques de ce document $Q_A$, et nous apparions $Q_N$ avec la question automatique $\hat{Q}_A$ donnant la similarité maximale tel que : $\hat{Q}_A = \underset{Q_A}{\operatorname{argmax}}\ \textit{sim}(Q_A, Q_N)$

Nous considérons que cet appariement est correct si $\hat{Q}_A$ a la même réponse que $Q_N$, et nous le notons $\mathrm{correct}(Q_A, Q_N)$. Cela signifie qu'ils peuvent être générés par le même triplet $(F, E, C)$ tel que : $\{\hat{Q}_A, Q_N\} \in \textit{Questions}(F, E, C)$. La réponse à ces deux questions se trouve ainsi être le support dans le texte pour l'élément $E$. Le tableau 2 recense les statistiques du corpus utilisé pour l'étude. La dernière colonne représente le nombre moyen d'appariements corrects pour une question naturelle donnée (c'est à dire le nombre moyen de questions automatiques ayant la même réponse), alors que l'avant dernière colonne représente le nombre moyen de candidats à l'appariement.

| collection | nb docs | nb questions naturelles ($Q_N$) | nb moyen $Q_N$ par doc | nb questions automatiques ($Q_A$) | nb moyen $Q_A$ par doc | nb moyen appariements corrects par $Q_N$ |
|---|---|---|---|---|---|---|
| V_antiq | 61 | 274 | 4.5 | 4672 | 76.6 | 4.2 |
| WP_arch | 96 | 302 | 2.4 | 36259 | 377.7 | 4.1 |
| CT_1GM | 16 | 241 | 15.1 | 7502 | 468.9 | 2.5 |
| WP_1GM | 123 | 319 | 2.6 | 50971 | 414.4 | 5.1 |
| **total** | **296** | **1136** | **3.8** | **99404** | **335.8** | **4.1** |

TABLE 2 – Description du corpus CALOR-QUEST

## 4.1 Appariement des questions automatiques et naturelles

Nous étudions comparativement 3 mesures de similarité textuelle $\textit{sim}(Q_A, Q_N)$ :

— le cosinus entre sacs de mots pondérés, $cos(Q_N, Q_A)$, qui reste une référence solide dans la famille des mesures basées sur les représentations creuses ;

— le cosinus entre des plongements des questions, $wavg-w2v(Q_N, Q_A)$, ces plongements étant simplement obtenus par la moyenne pondérée des plongements des mots qui les composent ;

— la similarité soft-cosinus $cos_M(Q_N, Q_A)$, qui consiste à introduire dans la formule du cosinus en sacs de mots une matrice de relations entre les mots, calculées à partir des plongements lexicaux (ou *embeddings*) de ces derniers (Charlet & Damnati, 2017).

Plus précisément, cette dernière similarité est donnée par :

$$cos_M(Q_N, Q_A) = \frac{Q_N^t.M.Q_A}{\sqrt{Q_N^t.M.Q_N}\sqrt{Q_A^t.M.Q_A}} \quad (1)$$

où $Q_N$ (resp. $Q_A$) représente le vecteur de sacs de mots pondérés de la question $Q_N$ (resp. $Q_A$) et $M$ la matrice dont l'élément $m_{i,j}$ exprime la relation entre les mots $i$ et $j$. Ici, elle est égale au carré de la similarité cosinus entre les plongements des mots $i$ et $j$. Les questions sont segmentées en mots, et subissent un prétraitement minimal (suppression des majuscules). Les poids des mots sont les TFIDF où les IDF sont estimés par collection, sur le corpus des questions automatiques.

Dans le cas particulier des questions, les pronoms interrogatifs et les prépositions jouent un rôle très structurant pour analyser finement le sens d'une question. Nous évaluons ainsi l'influence de la conservation des mots creux (liste *NLTK*) ou non dans les représentations de questions, ainsi qu'une variante du soft-cosinus, $cos_{M'}$, qui consiste à ne pas considérer de relations pour les mots creux.

## 4.2 Evaluation

Les performances d'appariement sont présentées dans le tableau 3. Pour cette évaluation, seul l'appariement de meilleur score est considéré. On constate tout d'abord que les performances sont liées de façon monotone au nombre de $Q_A$ candidates pour chaque $Q_N$ ($cf$ tableau 2). Plus ce nombre est élevé, moins les performances sont bonnes. C'est pourquoi l'on observe les meilleures performances sur le corpus V_antiq dont les documents sont plus courts et le nombre moyen de candidats à l'appariement est moins élevé (76.6 $Q_A$ par document en moyenne) alors que les performances les moins bonnes sont obtenues sur le corpus CT_1GM où le nombre de candidats par document est de 468.9 en moyenne. Le maintien des mots-creux améliore nettement les performances dans le cas du cosinus entre sacs de mots tandis que les performances de $cos_M$ sont dégradées et que leur influence est variable pour $wavg-w2v$. Les mots-creux en commun dans les questions favorisent l'appariement lorsqu'ils sont utilisés de façon stricte, mais la prise en compte des plongements de ces mots (que ce soit à travers la moyenne des plongements ou le soft-cosinus) ajoute du bruit. Les relations sémantiques sont néanmoins bénéfiques pour les mots pleins, ce qui est confirmé par les bonnes performances obtenues avec la version $cos_{M'}$. Une analyse des erreurs a révélé qu'une majorité des erreurs conduisent à apparier une $Q_N$ à une $Q_A$ issue d'une Frame différente (54% des erreurs sur V_antiq, 44% sur WP_arch, 61% sur CT_1GM et même 71% sur WP_1GM). Les erreurs résiduelles concernent majoritairement des erreurs vers une question issue de la même Frame mais portant sur un autre *FE*. Une analyse sémantique préalable des questions ou une prédiction du type de *FE* attendu pourrait améliorer la précision de l'appariement.

| collection | $cos$ | | $wavg-w2v$ | | $cos_M$ | | $cos_{M'}$ |
|---|---|---|---|---|---|---|---|
| *mots creux* | *avec* | *sans* | *avec* | *sans* | *avec* | *sans* | *avec* |
| V_antiq | 86.5 | 80.3 | 76.6 | 75.6 | 86.1 | 87.6 | 90.5 |
| WP_arch | 72.9 | 68.2 | 64.2 | 69.2 | 74.5 | 75.2 | 78.5 |
| CT_1GM | 66.8 | 64.7 | 61.0 | 66.4 | 70.5 | 72.2 | 74.7 |
| WP_1GM | 74.6 | 70.9 | 65.2 | 64.6 | 73.0 | 73.7 | 76.5 |

TABLE 3 – Pourcentage d'appariements corrects, selon la métrique de similarité utilisée

# 5 Conclusions

Nous avons proposé une méthode de génération semi-automatique de questions à partir d'un corpus analysé en cadres sémantiques avec l'objectif de produire un corpus pour la compréhension automatique de texte pour la langue française. Une première validation est proposée consistant à vérifier les performances d'un appariement par similarité textuelle entre les questions générées automatiquement et des questions naturelles produites par des annotateurs. Les bonnes performances obtenues ouvrent de nombreuses perspectives. En particulier, la représentation sémantique explicite des questions obtenues, inhérente à la méthodologie de construction, permettra d'envisager des méthodes d'appariement dépassant le cadre habituel de la représentation en sacs de mots et de mesurer l'apport de telles représentations. Enfin, l'apprentissage de modèles de compréhension de texte à partir des questions générées automatiquement afin d'en vérifier la validité sur des questions naturelles sera également une piste de recherche intéressante. L'ensemble des données collectées et générées sera mis à disposition de la communauté scientifique.

# Références

BAKER C. F., FILLMORE C. J. & LOWE J. B. (1998). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, p. 86–90 : Association for Computational Linguistics.

BÉCHET F., DAMNATI G., HEINECKE J., MARZINOTTO G. & NASR A. (2017). CALOR-Frame : un corpus de textes encyclopédiques annoté en cadres sémantiques. In *ACor4French – Les corpus annotés du français - Atelier TALN*, Orléans, France.

BOYER K. E. & PIWEK P. (2010). *Proceedings of QG2010 : The Third Workshop on Question Generation*. questiongeneration. org.

CHARLET D. & DAMNATI G. (2017). Simbow at semeval-2017 task 3 : Soft-cosine semantic similarity between questions for community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, p. 315–319.

CLARK P., COWHEY I., ETZIONI O., KHOT T., SABHARWAL A., SCHOENICK C. & TAFJORD O. (2018). Think you have solved question answering ? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv :1803.05457*.

DONG X., HONG Y., CHEN X., LI W., ZHANG M. & ZHU Q. (2018). Neural question generation with semantics of question type. In *CCF International Conference on Natural Language Processing and Chinese Computing*, p. 213–223 : Springer.

DUNN M., SAGUN L., HIGGINS M., GUNEY V. U., CIRIK V. & CHO K. (2017). Searchqa : A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv :1704.05179*.

FLOR M. & RIORDAN B. (2018). A semantic role-based approach to open-domain automatic question generation. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, p. 254–263.

HERMANN K. M., KOČISKÝ T., GREFENSTETTE E., ESPEHOLT L., KAY W., SULEYMAN M. & BLUNSOM P. (2015). Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, p. 1693–1701, Cambridge, MA, USA : MIT Press.

KOČISKÝ T., SCHWARZ J., BLUNSOM P., DYER C., HERMANN K. M., MELIS G. & GREFENSTETTE E. (2018). The narrative qa reading comprehension challenge. *Transactions of the Association of Computational Linguistics*, **6**, 317–328.

MAZIDI K. & NIELSEN R. D. (2014). Linguistic considerations in automatic question generation. In *ACL*.

NGUYEN T., ROSENBERG M., SONG X., GAO J., TIWARY S., MAJUMDER R. & DENG L. (2016). Ms marco : A human generated machine reading comprehension dataset. *arXiv preprint arXiv :1611.09268*.

PILLAI L. R., VEENA G. & GUPTA D. (2018). A combined approach using semantic role labelling and word sense disambiguation for question generation and answer extraction. In *2018 Second International Conference on Advances in Electronics, Computers and Communications (ICAECC)*, p. 1–6 : IEEE.

RAJPURKAR P., ZHANG J., LOPYREV K. & LIANG P. (2016). Squad : 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 2383–2392 : Association for Computational Linguistics.

SEO M., KEMBHAVI A., FARHADI A. & HAJISHIRZI H. (2016). Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv :1611.01603*.

TRISCHLER A., WANG T., YUAN X., HARRIS J., SORDONI A., BACHMAN P. & SULEMAN K. (2016). Newsqa : A machine comprehension dataset. *arXiv preprint arXiv :1611.09830*.

WANG S. & JIANG J. (2016). Machine comprehension using match-lstm and answer pointer. *arXiv preprint arXiv :1608.07905*.

WANG T., YUAN X. & TRISCHLER A. (2017). A joint model for question answering and question generation. *CoRR*, **abs/1706.01450**.

YUAN X., WANG T., GULCEHRE C., SORDONI A., BACHMAN P., ZHANG S., SUBRAMANIAN S. & TRISCHLER A. (2017). Machine comprehension by text-to-text neural question generation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, p. 15–25 : Association for Computational Linguistics.

ZHANG S., LIU X., LIU J., GAO J., DUH K. & VAN DURME B. (2018). Record : Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint arXiv :1810.12885*.

# Chunker différents types de discours oraux : défis pour l'apprentissage automatique

Iris Eshkol-Taravella[1, 2] Mariame Maarouf[2,3] Marie Skrovec[2] Flora Badin[2]
(1) MoDyCo UMR7114, 200 Avenue de la République, 92001 Nanterre, France
(2) LLL UMR7270, 10 Rue de Tours, 45065 Orléans, France
(3) Lattice UMR8094, 1 rue Maurice Arnoux, 92120 Montrouge, France
ieshkolt@parisnanterre.fr, maarouf.mariame@gmail.com, marie.skrovec@univ-orleans.fr, flora.badin@univ-orleans.fr

RÉSUMÉ

Le travail décrit le développement d'un chunker pour l'oral par apprentissage supervisé avec les CRFs, à partir d'un corpus de référence de petite taille et composé de productions de nature différente : monologue préparé vs discussion spontanée. La méthodologie respecte les spécificités des données traitées. L'apprentissage tient compte des résultats proposés par différents étiqueteurs morpho-syntaxiques disponibles sans correction manuelle de leurs résultats. Les expériences montrent que le genre de discours (monologue vs discussion), la nature de discours (spontané vs préparé) et la taille du corpus peuvent influencer les résultats de l'apprentissage, ce qui confirme que la nature des données traitées est à prendre en considération dans l'interprétation des résultats.

ABSTRACT

**Chunking different spoken speech types : challenges for machine learning**

This paper describes the development of a chunker for spoken data by supervised machine learning using the CRFs, based on a small reference corpus composed of two discourse types: prepared monologue vs spontaneous talk in interaction. The methodology respects the specific character of the processed data. The machine learning considers the results of several available taggers, without manual correction of their results. Experiments show that the discourse type (monologue vs free talk), the speech nature (spontaneous vs prepared) and the corpus size can influence the results of the machine learning process. The type of data should therefore be considered in interpreting the results.

MOTS-CLÉS : chunking, apprentissage supervisé, CRF, segmentation automatique de l'oral, corpus oral, variation discursive, genre

KEYWORDS: chunking, machine learning, CRF, automatic segmentation of oral data, oral corpus, discursive variation, genre

# 1 Introduction

La notion de phrase étant généralement considérée comme peu pertinente pour l'analyse et le traitement de l'oral (Blanche-Benveniste et al, 1990 ; Groupe de Fribourg, 2012), différents types d'unités de segmentation de l'oral ont été proposés par des recherches menées dans le cadre de projets comme Rhapsodie ou Orfeo. Le projet SegCor[1] porte aussi sur la segmentation des corpus oraux et propose une segmentation multiniveau. Son premier niveau est une segmentation en unités minimales syntaxiques en termes de constituance, appelées chunks.

Les chunks sont des constituants continus et non récursifs (Abney 1991). Le chunking identifie la structure syntaxique superficielle d'un énoncé et peut être effectué automatiquement. Il est fondé sur un étiquetage morphosyntaxique préalablement réalisé. Ce type d'annotation est particulièrement pertinent pour l'oral car le discours oral est parfois composé d'énoncés non « finalisés » ce qui rend une analyse syntaxique complète difficile. Blanche-Benveniste (1997) a démontré que ces constituants sont le lieu de réalisation privilégié des réparations à l'oral.

Plusieurs stratégies sont possibles pour développer un chunker. Les méthodes symboliques ont été testées dans le cadre des travaux de (Blanc et *al*, 2008, 2010, Antoine et *al*, 2008) où des cascades de transducteurs développées chunkent des transcriptions de l'oral en se fondant sur des ressources lexicales et syntaxiques. L'apprentissage automatique supervisé semble être particulièrement performant sur cette tâche comme montrent les recherches de (Sha et Pereira, 2003, Tellier et *al*, 2012, 2014, Tsuruoka et *al*, 2009). Dans la suite du travail de (Tellier et *al*, 2014), la recherche présentée ici utilise la méthode de l'apprentissage supervisé. Les productions orales se caractérisent par une grande variété discursive : variété situationnelle (conversation privée, débat public…), tâches langagières (expliquer, raconter, décrire…), genres (récits de voyage, interview…) ou registre de langue (courant, familier, soutenu). La nature des données traitées influence et guide le processus d'apprentissage. Dans le travail de (Tellier et *al*, 2014), le corpus de référence était composé d'entretiens sociolinguistiques ; dans celui-ci nous nous fondons sur deux autres situations de communication : une conférence et une discussion spontanée entre plusieurs personnes lors d'un repas.

# 2 Constitution du corpus de référence

Les données traitées proviennent de deux grands corpus de français parlé contemporain : ESLO2[2] et CLAPI[3]. Deux types de discours sont sélectionnés : une conférence donnée par un locuteur, un monologue préparé (10 minutes, 2120 tokens) dans le corpus ESLO2 (M) et une discussion entre trois personnes, une interaction spontanée, se déroulant dans un contexte d'ordre privé (10 minutes, 2461 tokens) dans le corpus CLAPI (R).

1 Un projet franco-allemand, financé par l'Agence Nationale de Recherche (ANR-15-FRAL-0004)

2 Enquêtes Sociolinguistiques à Orléans, http://eslo.huma-num.fr/

3 Corpus de LAngue Parlée et Interaction, http://clapi.ish-lyon.cnrs.fr/

## 2.1 Prétraitement

Les deux fichiers utilisés pour ce travail sont prétraités en termes de segmentation et d'annotation. Les tokens, les annotations et le signal sonore sont d'abord alignés semi-automatiquement[4].
Les unités polylexicales (*comme ça*, *plein de*, *ciné club*) sont repérées ensuite grâce à la ressource Lefff (Sagot et al., 2010). Le résultat du prétraitement est montré dans la Figure 1.

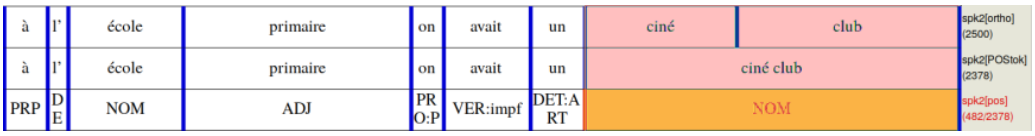



Figure 1 : Résultat et visualisation du prétraitement sous Praat[5]

## 2.2 Typologie des chunks

La typologie de chunks est fondée sur celle présentée dans Tellier *et al.* (2014) et complétée par deux nouvelles étiquettes (FNO et ARTIC). Elle contient neuf catégories :

– adjectival phrase (AP) : chunk adjectival - l'adjectif tête placé après le verbe (*elle est trop jolie*) ;
– adverbial phrase (AdP) : chunk adverbial - un syntagme dont la tête est un adverbe (*peut-être*) ;
– nominal phrase (NP) : chunk nominal - les syntagmes nominaux intégrant les adjectifs placés avant et après le nom et les pronoms non clitiques (*tes belles chaussures*) ;
– prepositional phrase (PP) : chunk prépositionnel - les syntagmes introduits par une préposition (*de loin*) ;
– verbal phrase ou verbal nucleus (VP) : chunk verbal – les syntagmes organisés autour d'une tête verbale, associée à ses clitiques (*on nous entend*), fléchie ou non ;
– ponctuation (SENT) : les transcriptions ne contiennent pas de marques typographiques, sauf des points d'interrogation conservés pour plus de lisibilité ;
– articulateur (ARTIC) : une catégorie qui regroupe des éléments non autonomes reliant des unités de différents niveaux, qu'il y ait dépendance syntaxique ou non, comme les pronoms relatifs, les conjonctions, les marqueurs discursifs (*et, que, lequel, enfin, mais, du coup,* etc.);
– forme noyau (FNO) : inspirée des travaux de Benzitoun *et al.* (2012), cette catégorie regroupe des éléments autonomes, non périphériques, constituant à eux seuls une unité illocutoire (*oui, ouf, merde, d'accord, voilà, bonjour, salut, mince, santé*, etc.) ;
– inconnu (UNKNOWN) : une catégorie regroupant les chunks non identifiés, comme les amorces de mots, les mots mal orthographiés, etc.

4 Découpage en unités poylexicales et annotation en POS : Treetagger (Schmid, 1994), Dismo (Christodloulides et al., 2014) et réalignement manuel sur le signal sonore

5 Praat est un outil de transcription et d'annotation manuelle de l'oral (http://www.fon.hum.uva.nl/paul/praat.html).

## 2.3 Annotation manuelle

Les deux corpus prétraités sont d'abord annotés par deux chercheurs selon la typologie établie. L'accord inter-annotateur calculé en appliquant le kappa de Cohen (Cohen, 1960) est de 88%. La troisième annotation de consensus est effectuée par la suite sur le même corpus, elle sert de corpus de référence et d'évaluation pour l'apprentissage automatique. L'annotation est réalisée à l'aide du logiciel Praat (Boersma et Van Heuven, 2001) et en utilisant le format BILOU[6] (Ratinov et Roth, 2009) permettant de délimiter une unité mais aussi de déterminer la place de chaque terme au sein de cette unité. Grâce à Praat, les annotateurs ont accès à l'écoute des enregistrements pour mieux comprendre certaines situations ce qui n'a pas été effectué dans le cadre du travail de (Tellier *et al*. 2014). Le corpus ainsi annoté contient 1069 chunks dans M et 1455 chunks dans R répartis de manière hétérogène dans les deux corpus (la présence importante de PP 30% dans M vs 11% dans R contrairement au VP représentant 40% dans R vs 23% dans M, etc.).

# 3 Apprentissage automatique

L'apprentissage automatique vise à indiquer des frontières de chaque chunk, mais aussi à déterminer son type. Le corpus de référence ayant une petite taille, nous optons pour le modèle des CRFs (*Conditional Random Fields*) linéaires (Lafferty *et al*, 2001) qui a déjà fait preuve d'une bonne performance pour cette tâche (Sha et Pereira, 2003, Tellier et *al*, 2012, 2014, Tsuruoka et *al*, 2009). Le chunking s'applique sur le corpus étiqueté en POS. Tellier *et al.* (2014) ont montré qu'on peut apprendre un chunker propre à l'oral avec des POS non corrigées et avec un corpus de référence de petite taille. Les auteurs arrivent à 88% de micro-average. Nous poursuivons la même démarche mais avec une méthodologie redéfinie en fonction de la spécificité des données orales : (1) les données traitées sont plus hétérogènes car elles comprennent deux types de discours oral ; (2) les annotateurs humains ont systématiquement recours à l'écoute du son pour déterminer les choix d'annotation ; (3) le jeu d'étiquettes est retravaillé (l'ajout de deux nouvelles étiquettes *ARTIC* et *FNO*); (4) les résultats de plusieurs étiquetages morpho-syntaxiques sont ajoutés dans les traits intégrés au modèle CRF ce qui permet de vérifier si une série d'étiquettes POS non corrigées proposées par différents étiqueteurs pour le même mot améliore les résultats du chunking et quels outils parmi ceux testés sont les plus pertinents pour le corpus oral traité.

Quatre étiqueteurs sont testés : TreeTagger (Schmidt, 1994) ; SEM (Tellier et al., 2012) exploité par (Tellier et al., 2014) et utilisant les étiquettes morpho-syntaxiques de (Crabbé et al., 2008) ; parseur en dépendance syntaxique (Kahane et *al.*, 2017) développé dans le cadre du projet Orfeo, d'où nous extrayons uniquement les POS et les POS du gouverneur syntaxique du token courant ; Perceo (Benzitoun et *al.*, 2012), étiqueteur adapté pour l'oral qui a la particularité de posséder une étiquette FNO, étiquette aussi présente dans notre typologie de chunks.

6 B pour Begin, premier token du chunk; I pour In, un élément à l'intérieur d'un chunk ; L pour Last, dernier élément du chunk ; O pour Out, un élément extérieur, absent dans le corpus car tous les tokens font partie d'un chunk ; U pour Unit, un chunk composé d'un seul token.

Les expériences sont effectuées sur trois corpus : ESLO2 (M), CLAPI (R), ESLO2+CLAPI (M+R). L'objectif est de vérifier si le genre de discours (monologue/discussion entre 3 personnes), la nature de discours (spontané/préparé) et la taille du corpus peuvent influencer les résultats d'apprentissage. De nombreuses configurations sont testées afin d'obtenir des résultats exhaustifs en combinant et variant le nombre de patrons [token + POS] pris en compte[7]. Pour le parseur d'Orfeo, deux combinaisons supplémentaires sont testées (1) en prenant en compte uniquement l'étiquette POS du token, (2) l'étiquette POS du token et de son gouverneur. En premier lieu, pour chacune des combinaisons la prise en compte du token de la ligne courante est testée. Ensuite, les trois combinaisons donnant les meilleurs résultats pour chaque corpus sont sélectionnées pour les tests en incluant token+1 et token-1. Après isolation du meilleur résultat, d'autres colonnes sont ajoutées comme par exemple le lemme, récupéré depuis l'annotation TreeTagger, pour tester une possible amélioration du score. La Figure 2 montre les meilleures combinaisons de patrons pour chaque corpus.

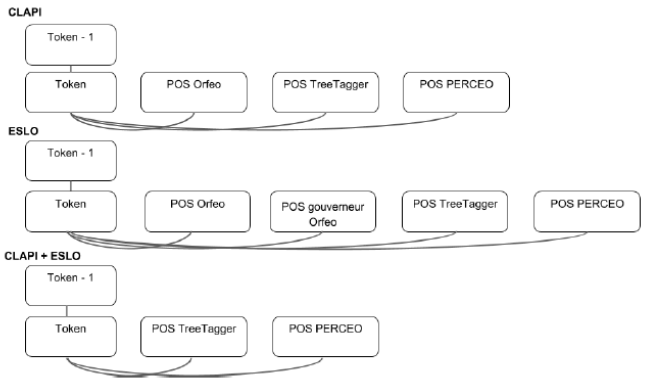



Figure 2 : Meilleures combinaisons de patrons pour chaque corpus

# 4 Résultats et évaluation

L'évaluation des résultats est effectuée en validation croisée à 10 plis[8] et sur trois corpus (M, R, M+R) séparément. Trois mesures différentes sont utilisées pour évaluer les performances de l'annotation : la précision, le rappel et la F-mesure[9]. Ces mesures sont calculées pour chaque type de chunks et à partir de ces résultats, on obtient la micro-average[10]. Cette pondération permet

7 token+SEM, token+SEM+TTG, token+SEM+TTG+Orfeo, token+SEM+TTG+Orfeo+Perceo, token+Orfeo, token+Orfeo+TTG..., token+1 et token-1.

8 En réalisant un apprentissage sur 9/10 des exemples, on se prive de peu de données tout en s'assurant de fournir une évaluation peu « biaisée » car elle est une moyenne de plusieurs expériences.

9 la moyenne harmonique du rappel et de la précision

10 la moyenne pondérée des résultats obtenus des différents types de chunks

d'attribuer plus ou moins de poids aux résultats en fonction de leur taux de présence dans le corpus. Ainsi, plus une catégorie de chunks est présente dans le corpus, plus son score aura d'importance dans le calcul de la micro-average et inversement.

| | M | R | M+R |
|---|---|---|---|
| | 85,8% | 83,2% | 85,7% |
| POS TreeTagger | x | x | x |
| POS Perceo | x | x | x |
| POS Orfeo | x | x | |
| Gouv Orfeo | x | | |
| Tok_courant Tok_precedent | x | x | x |

Tableau 1 : Tableau de meilleurs résultats obtenus en termes de micro-average

Pour le corpus R, les meilleurs résultats sont produits par la combinaison qui regroupe les outils TreeTagger, PERCEO et le parseur d'Orfeo auxquels s'ajoutent le token courant et le token précédent (83,2 %). Les meilleurs scores obtenus pour le corpus M proviennent de l'application d'une combinaison similaire à celle de R sauf qu'en plus de l'étiquette POS récupérée par le parsing en dépendance du parseur d'Orfeo, l'étiquette du gouverneur est aussi présente (85,8 %). Le monologue, extrait d'une conférence, est composé d'énoncés plus longs et ne contient pas d'interaction, les liens de dépendances y sont plus présents. Dans le cas du corpus M+R, les meilleurs résultats sont obtenus en n'utilisant que les POS de TreeTagger et de Perceo, en plus du token précédent au token courant (85,7 %). Ces résultats montrent que l'utilisation des étiquettes proposées par deux outils de l'oral, PERCEO et le parseur d'Orfeo, dans les patrons est tout à fait pertinente. La taille du corpus est un critère important aussi car le corpus plus long n'a pas besoin de beaucoup d'étiquettes POS et peut se contenter des résultats de deux outils. Il est étonnant que TreeTagger semble être plus pertinent dans ce cas que le parseur d'Orfeo. Le corpus M, un monologue préparé, se prête mieux à l'apprentissage que le corpus R, la discussion spontanée.

L'évaluation du chunker par étiquette montre que l'étiquette FNO obtient les moins bons résultats (23,52% de F-mesure). En effet, certains tokens sont ambigus, comme par exemple *oui* , *ouais*, *non*, tantôt FNO, tantôt articulateurs (marqueurs discursifs). Ainsi, un *ouais* en réponse à une question sera considéré comme prédicat autonome (un « mot-phrase ») et donc annoté (FNO), comme ici :

ELI je [VP B] vous [VP I] sers [VP L] ?
BEA ouais [FNO U]

En revanche cette même forme peut être considérée comme élément périphérique au prédicat. Il s'agira alors d'un articulateur discursif non autonome, comme dans l'exemple ci-dessous, où *ouais* opère comme balise de clôture du tour de ELI :

ELI non [ARTIC U] mais [ARTIC U] tu [VP B] sais [VP L]
tu [VP B] en [VP I] mets [VP L] pas [AdP B] beaucoup [AdP L]
tu [VP B] en [VP I] mets [VP L] un [NP B] fond [NP L] ouais [ARTIC U]

On relève quelques autres erreurs courantes. Ainsi, de nombreux chunks NP sont annotés comme PP à cause de l'ambiguïté entre la préposition *de* suivi du déterminant défini et l'article partitif (*du, de la, etc.*), tous les deux ayant la même forme. Par ailleurs, un quart des AP sont considérés comme des VP car souvent un token de type AP suit un chunk de type VP. D'une manière générale, les frontières de chunks (les étiquettes B, L, U) sont mieux annotées (Tableau 2).

| | B | I | L | U |
|---|---|---|---|---|
| R | 0,94 | 0,86 | 0,91 | 0,94 |
| M | 0,92 | 0,87 | 0,93 | 0,9 |
| M+R | 0,93 | 0,86 | 0,92 | 0,93 |

Tableau 2 : Résultats de F-mesure pour les étiquettes BILU

# 5 Conclusion

Les productions orales se caractérisent par une grande variété discursive. L'article décrit le développement d'un chunker par apprentissage automatique avec les CRFs en utilisant un corpus de référence de petite taille comprenant les données orales de nature différente : monologue dans le cadre d'une conférence vs discussion spontanée entre 3 personnes lors d'un repas. Un genre et un type de discours peuvent influencer les résultats d'apprentissage. Ainsi, les résultats du parsing en dépendance sont plus pertinents à intégrer au modèle CRF pour le monologue où les énoncés longs se prêtent plus à ce type d'analyse. Les FNO obtiennent de meilleurs scores dans une discussion car ils y sont plus nombreux. La nature des données traitées est donc à prendre en considération dans l'interprétation des résultats. Plusieurs perspectives sont envisagées : (1) d'ajouter certaines informations issues des enregistrements comme la prosodie ; (2) de laisser les deux options dans les cas où les annotateurs humains hésitent entre différentes étiquettes possibles ce qui améliorera les résultats du chunker ; (3) d'ajouter des règles d'annotation pour certains phénomènes récurrents et systématiques comme la précision qu'un tour de parole commence toujours par une frontière B ou U ; (4) d'intégrer dans le corpus d'apprentissage le maximum de situations de communication pour généraliser le développement d'un chunker pour l'oral.

# Remerciements

Ce travail a été effectué dans le cadre du stage de Mariame Maarouf, co-encadré par Isabelle Tellier, qui nous a quittés le 1 juin 2018. Nous tenons ici à lui rendre un hommage affectueux et à lui témoigner notre gratitude pour son enthousiasme, ses idées et ses conseils avisés, sans lesquels cet article n'aurait pu voir le jour.

# Références

Abney S. (1991). *Parsing by chunks*. In R. Berwick, R. Abney, and C. Tenny, editors, Principle-based Parsing. Kluwer Academic Publisher.

Antoine J.-Y., Mokrane A., Friburger N. (2008). Automatic rich annotation of large corpus of conversational transcribed corpus. Actes de *LREC 2008*.

Benzitoun C., Fort K., Sagot B. (2012). TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe. Actes de *JEP-TALN 2012,* 99-112.

Blanc O., Constant M., Dister A., Watrin P. (2008). Corpus oraux et chunking. Actes de *Journées d'étude sur la parole (JEP).*

Blanc O., Constant M., Dister A., Watrin P. (2010). Partial parsing of spontaneous spoken french. Actes de *LREC'10.*

Blanche-Benveniste C., Bilger M., Rouget C., Van Den Eynde K. (1990). *Le français parlé*. Études grammaticales, Paris, CNRS Éditions.

Blanche-Benveniste C. (1997). *Approches de la langue parlée en français*. Paris, Ophrys.

Boersma P., Van Heuven V. (2001). Speak and unSpeak with Praat. *Glot International,* 5(9/10), 341-347.

Christodoulides G., Avanzi M., Goldman J-P. (2014). DisMo: A Morphosyntactic, Disfluency and Multi-Word Unit Annotator. Actes de *LREC'14*.

Cohen J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.

Groupe de Fribourg, (2012), *Grammaire de la période*, Berne, Peter Lang.

Kahane S., Deulofeu J., Gerdes K., Nasr A., Valli A. (2017). Annotation micro et macrosyntaxique manuelles et automatique de français parlé. *Journée Floral*, mars 2017, Orléans.

Lafferty J., McCallum A., Pereira F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. Actes de *ICML 2001,* 282-289.

Ratinov L., Roth D. (2009). Design challenges and misconceptions in named entity recognition. Actes de *CoNLL*.

Sagot B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. Actes de *LREC 2010*.

Schmidt H. (1994). Probabilistic part-of-speech tagging using decisions trees. Actes d'*International Conference on New Methods in Language Processing*, 44-49.

Sha F., Pereira F. (2003). Shallow parsing with conditional random fields. Actes de *HLT-NAACL 2003*, 213-220.

Tellier I., Duchier D., Eshkol I., Courmet A., Martinet M. (2012). Apprentissage automatique d'un chunker pour le français, Actes de *TALN 2012*.

Tellier I., Eshkol-Taravella I., Dupont Y., Wang I. (2014). Peut-on bien chunker avec de mauvaises étiquettes pos ? Actes de *TALN2014*.

Tsuruoka Y., Tsujii J., Ananiadou S. (2009). Fast full parsing by linear-chain conditional random fields. Actes de *EACL 2009*.

# Classification automatique des procédés de traduction

Yuming Zhai Gabriel Illouz Anne Vilnat
LIMSI-CNRS, Univ. Paris-Sud, Univ. Paris-Saclay, France
{prénom.nom}@limsi.fr

RÉSUMÉ

En vue de distinguer la traduction littérale des autres procédés de traduction, des traducteurs et linguistes ont proposé plusieurs typologies pour caractériser les différents procédés de traduction, tels que l'équivalence idiomatique, la généralisation, la particularisation, la modulation sémantique, etc. En revanche, les techniques d'extraction de paraphrases à partir de corpus parallèles bilingues n'ont pas exploité ces informations. Dans ce travail, nous proposons une classification automatique des procédés de traduction en nous basant sur des exemples annotés manuellement dans un corpus parallèle (anglais-français) de *TED Talks*. Même si le jeu de données est petit, les résultats expérimentaux sont encourageants, et les expériences montrent la direction à suivre dans les futurs travaux.

ABSTRACT

**Automatic Classification of Translation Processes**

In order to distinguish literal translation from other translation processes, translators and linguists have proposed several typologies to characterize different translation processes, such as idiomatic equivalence, generalization, particularization, semantic modulation, etc. However, the techniques to extract paraphrases from bilingual parallel corpora have not exploited this information. In this work, we propose an automatic classification of translation processes, based on manually annotated examples in an English-French parallel corpus of *TED Talks*. Even with a small dataset, the experimental results are encouraging and our experiments show the direction to follow in future work.

MOTS-CLÉS : procédés de traduction, classification automatique, extraction de paraphrases.

KEYWORDS: translation processes, automatic classification, paraphrase extraction.

## 1 Introduction

Les procédés de traduction sont étudiés depuis longtemps (Vinay & Darbelnet, 1958; Chuquet & Paillard, 1989; Molina & Hurtado Albir, 2002). Ils distinguent les traductions littérales des autres procédés de traduction au niveau sous-phrastique. Prenons comme exemple ces deux traductions humaines non littérales : la première traduction préserve exactement le sens, où l'expression figée *à la hauteur de* possède un sens figuré *« capable de résoudre »* ; en revanche, la deuxième traduction est plus compliquée, où il existe une inférence textuelle entre le segment source et la traduction.

*(1.EN) a solution* ***that's big enough to solve*** *our problems*
*(1.FR) une solution* ***à la hauteur de*** *nos problèmes*
*(2.EN) and* ***that scar has stayed with him*** *for his entire life*
*(2.FR) et que, toute sa vie,* ***il a souffert de ce traumatisme***

Les traductions non littérales peuvent poser des difficultés pour l'alignement de mots automatique (Dorr *et al.*, 2002; Deng & Xue, 2017), ou causer des changements de sens dans certains cas. Cependant, à notre connaissance, les techniques de traitement automatique des langues n'ont pas explicitement exploité ces procédés de traduction. Bannard & Callison-Burch (2005) ont proposé d'exploiter les techniques de traduction automatique pour extraire des paraphrases à partir de corpus parallèles bilingues. Leur hypothèse est que deux segments monolingues sont des paraphrases potentielles s'ils partagent des traductions communes dans une autre langue. Actuellement, la plus grande ressource de paraphrases, PPDB (ParaPhrase DataBase) (Ganitkevitch *et al.*, 2013; Pavlick *et al.*, 2015b), a été construite selon cette méthode. En revanche, Pavlick *et al.* (2015a) ont révélé qu'il existe d'autres relations que l'équivalence stricte (paraphrase) dans PPDB (*i.e. Implication (dans les deux directions), Exclusion, Lié à* et *Indépendant*) [1]. Des traductions « pivots » non littérales dans des corpus parallèles bilingues peuvent influencer l'équivalence stricte entre les candidats de paraphrases extraits, néanmoins elles n'ont pas reçu assez d'attention pendant cette exploration de corpus.

De notre côté, en nous basant sur les théories développées en traduction, nous avons annoté et analysé manuellement un corpus parallèle anglais-français de *TED Talks*. Ce travail nous permet de proposer une typologie de procédés de traduction adaptée à notre corpus, ainsi que d'établir le guide d'annotation. Dans cet article, nous présentons une classification automatique des procédés de traduction en utilisant ce corpus annoté. Après avoir présenté les travaux précédents (section 2), nous décrivons l'annotation manuelle et le jeu de données (section 3). La section 4 présente les traits exploités pour la classification automatique et la section 5 montre les résultats et les analyse. La conclusion et les perspectives suivent dans la section 6.

# 2 Travaux précédents

Vinay & Darbelnet (1958) ont identifié des procédés de traduction directe et indirecte, ces derniers correspondent aux cas où une traduction littérale est inacceptable, ou lorsque les asymétries structurelles ou conceptuelles entre la langue source et la langue cible ne sont pas négligeables. Ces travaux ont été poursuivis par Newmark (1981, 1988) et Chuquet & Paillard (1989). Plus récemment, Molina & Hurtado Albir (2002) ont proposé leur propre classification basée sur une étude de la traduction des éléments culturels du roman *Cent ans de solitude* de l'espagnol vers l'arabe. Pour le couple anglais-chinois, Deng & Xue (2017) identifient, catégorisent et quantifient semi-automatiquement sept types de divergences de traduction, causées par des traductions non littérales ou des différences grammaticales inter-linguistiques [2]. Nous annotons le corpus selon une typologie inspirée par ces travaux précédents, mais aussi adaptée au corpus de *TED Talks*.

Récemment, différents modèles ont été proposés pour détecter automatiquement des divergences en traduction dans des corpus parallèles. Le but est de filtrer automatiquement des couples de phrases divergents afin d'améliorer la performance des systèmes de traduction automatique. Carpuat *et al.* (2017) ont introduit un détecteur de divergence cross-lingue basé sur SVM, en utilisant des traits en alignement de mots et en longueur de phrase. Vyas *et al.* (2018) ont proposé une approche basée sur des réseaux neuronaux profonds, et l'entraînement ne demande pas d'annotation manuelle. D'une façon non supervisée, Pham *et al.* (2018) ont généré des plongements phrastiques en fonction de

1. Exclusion : X est le contraire de Y ; X et Y s'excluent mutuellement. Lié à : X est lié d'une certaine manière à Y. (*p. ex. pays / patriotique*). Indépendant : X n'est pas lié à Y.

2. Encodage lexical ; différence de transitivité ; absence de mots grammaticaux spécifiques à une langue ; différence de catégories de phrases ; différence dans l'ordre de mots ; éléments omis ; paraphrases structurelles.

la similarité entre les mots. Ils mesurent l'équivalence sémantique entre les phrases afin de guider le filtrage. Contrairement à ces efforts qui ont lieu au niveau phrastique et effectuent une décision binaire, nous classifions automatiquement différents procédés de traduction au niveau sous-phrastique à partir d'exemples annotés manuellement. Ceci permettra d'identifier certains procédés de traduction qui peuvent provoquer des divergences sémantiques, tandis que d'autres conservent le sens original.

# 3 Annotation manuelle et description des données

Afin de modéliser les choix de traduction effectués par les traducteurs humains au niveau sous-phrastique, nous avons annoté un corpus parallèle trilingue (anglais-français, anglais-chinois) de *TED Talks* [3] en procédés de traduction (Zhai, 2018; Zhai *et al.*, 2018). Le corpus est composé de transcriptions et de traductions humaines de présentations orales. Nous choisissons ce genre spécifique parce que plusieurs domaines sont couverts dans les présentations, et la diversité des phénomènes de traduction est entre celle présente dans les corpus littéraires et techniques. L'accord inter-annotateur Kappa (Cohen, 1960) pour annoter le corpus de contrôle anglais-français et anglais-chinois est de 0,67 et 0,61, tous proches du seuil pour être suffisant [4]. Cela indique que la tâche de l'annotation manuelle est complexe.

Nous présentons dans la table 1 une brève définition, un exemple typique et le nombre d'instances pour chaque catégorie à classifier automatiquement pour le couple anglais-français [5]. Nous combinons *transposition* et *mod+trans* dans une catégorie *contient_transposition*, où la classe *modulation* est considérée comme neutre. Dans ce présent travail, nous menons des expériences dans un scénario simplifié, où nous connaissons déjà les frontières des couples bilingues, et nous ne prédisons que le procédé de traduction. Par exemple, étant donné le couple *deceptive* → *une illusion*, le but est de prédire son étiquette *contient_transposition*.

# 4 L'ingénierie des traits pour la classification automatique

Nous avons exploité quatre groupes de traits ci-dessous pour le couple anglais-français. Les jeux d'étiquettes des deux langues pour l'analyse morpho-syntaxique, l'analyse syntaxique en constituant et en dépendance ont été convertis en trois jeux unifiés et compacts (Petrov *et al.*, 2012).

**Analyse morpho-syntaxique (PoS)** 1) L'analyse est faite par *Stanford CoreNLP* (Manning *et al.*, 2014) pour les deux langues. Pour chaque langue, le nombre d'occurrence de chaque étiquette est compté dans un vecteur. Nous calculons aussi la similarité cosinus entre ces deux vecteurs (sur tous les mots et sur seulement les mots pleins [6]).
2) Nous vérifions le patron de changement de séquence de PoS selon une liste construite manuellement. Par exemple le couple *methodologically* → *de façon méthodologique* correspond au patron *ADV* → *ADP NOUN ADJ*.

3. `https://www.ted.com/`
4. La valeur minimum pour atteindre un accord suffisant est de 0,61.
5. Notez qu'il existe d'autres règles d'annotation détaillées dans le guide d'annotation.
6. Les étiquettes de mots pleins contiennent : ADJ, ADV, NOUN, PROPN, VERB. Si un segment ne contient aucun mot plein, nous utilisons le segment original.

| Procédé | Définition et exemple typique |
|---|---|
| littéral<br>(3771) | Traduction mot à mot.<br>*certain* ***kinds*** *of* → *certains* ***types*** *de* |
| équivalence<br>(289) | Traduction non littérale des proverbes ou des expressions figées ; Une traduction mot à mot est possible mais le traducteur exprime différemment, sans changer le sens ni les catégories grammaticales.<br>***back then*** → ***à l'époque*** |
| transposition<br>(289) | Modification des catégories grammaticales sans en changer le sens.<br>*unless* ***something changes*** → *à moins qu'****un changement ait lieu*** |
| modulation<br>(195) | Modulation métonymique et grammaticale (Chuquet & Paillard, 1989) ; Changement du point de vue ; Changement du sens possible.<br>***that scar has stayed with him*** → ***il a souffert de ce traumatisme*** |
| mod+trans<br>(53) | Combinaison des transformations de *Modulation* et de *Transposition*, ce qui peut rendre l'alignement de mots difficile.<br>*this is a completely* ***unsustainable*** *pattern* → *il est absolument* ***impossible de continuer sur*** *cette tendance* |
| généralisation<br>(86) | Plusieurs mots ou expressions sources peuvent être traduits en un mot ou une expression cible avec un sens plus général, le traducteur utilise ce dernier.<br>*as we* ***sit here*** *in ...* → *alors que nous* ***sommes*** *à ...* |
| particularisation<br>(215) | Le mot ou l'expression source peut être traduit en plusieurs mots ou expressions cibles avec un sens plus spécifique. Le traducteur en choisit un selon le contexte.<br>*they* ***have*** *a screen* → *ils* ***sont équipés d'****un écran* |

TABLE 1 – Définition, exemple typique et nombre d'instances des procédés de traduction à classifier automatiquement.

**Surface** 3) Le nombre de tokens dans les deux segments ($l_e$, $l_f$), le ratio de ces nombres ($l_e/l_f$, $l_f/l_e$), la distance Levenshtein (Levenshtein, 1966) entre les segments.

**Analyse syntaxique** 4) L'analyse syntaxique en constituant est faite par *Bonsai* (Candito *et al.*, 2010) pour le français, par *Stanford CoreNLP* pour l'anglais [7]. Nous comparons les étiquettes PoS pour un couple de mots ; les étiquettes du nœud non terminal pour un couple de segments ; la catégorie des étiquettes (*i.e.* verbe → syntagme verbal) pour un mot traduit par un segment ou vice versa.
5) L'analyse syntaxique en dépendance est faite par *Stanford CoreNLP* pour les deux langues afin de partager le même jeu d'étiquettes. À l'intérieur des segments, nous comptons le nombre d'occurrence de chaque relation de dépendance. À l'extérieur des segments, parmi les mots liés en dépendance dans chaque langue, nous gardons ceux qui sont manuellement alignés préalablement. Ensuite, nous comptons le nombre d'occurrence de chaque relation de dépendance que les mots à l'intérieur du segment entretiennent avec ces mots de contexte.

**Ressource externe** 6) Nous calculons la similarité cosinus entre les plongements provenant de *ConceptNet Numberbatch* (Speer *et al.*, 2017). Cette ressource est multilingue et le système basé sur *ConceptNet* a remporté la première place dans la tâche "Similarité sémantique lexicale multilingue et cross-lingue" de SemEval2017 (Camacho-Collados *et al.*, 2017; Speer & Lowry-Duda, 2017).

7. Pour l'analyse en constituant, *Bonsai* est beaucoup plus rapide que *Stanford CoreNLP* et a moins d'erreurs évidentes.

Certaines expressions multi-mots ont leur propre plongement dans cette ressource. Sinon, nous calculons la moyenne des plongements sur seulement les mots pleins. Les mêmes traits ont été calculés pour les segments lemmatisés [8].
7) La ressource *ConceptNet* fournit des assertions sous forme de triplet : un couple de mots ou expressions liés par une relation [9]. Dans cette ressource multilingue, nous vérifions si un couple anglais-français est directement lié ; indirectement lié par un autre segment français ou simplement pas lié. Trois formes sont testées : forme originale, forme lemmatisée et forme lemmatisée filtrée. [10]
8) Sur la forme lemmatisée filtrée, nous calculons le pourcentage des tokens bilingues qui sont liés avec une relation de dérivation, en basant sur la ressource *ConceptNet*. Par exemple *deceptive* et *illusion* ne sont pas directement liés dans la ressource, mais tous les deux sont liés à *illusoire*. Ainsi nous considérons qu'il existe un lien de dérivation entre eux.

**Alignement de mot** Pour ce groupe de traits, nous avons exploité la table de probabilité de traduction lexicale générée par l'outil statistique d'alignement de mots *Berkeley Word Aligner* (Liang *et al.*, 2006), entraîné sur un corpus parallèle anglais-français combiné de *TED Talks* et d'une partie du corpus Paracrawl [11] (au total 1.8M de couples de phrases et 41M de tokens anglais) :

9) L'entropie des distributions de probabilités de traduction lexicale (Gray, 1990; Carl & Schaeffer, 2017) : calculée selon cette équation : $H(X) = \sum_i P(x_i)I(x_i) = -\sum_i P(x_i)log_e P(x_i)$. Nous calculons l'entropie moyenne sur des mots pleins. Une entropie plus grande indique que les mots possèdent des sens plus généraux ou qu'ils sont polysémiques. Le même trait est calculé sur les mots pleins lemmatisés.
10) La pondération lexicale bidirectionnelle sur les mots pleins, en supposant un alignement de mots *n-m* ($a$) entre deux segments ($\bar{e}$ et $\bar{f}$). Selon l'équation proposée par Koehn *et al.* (2003) [12], pour calculer la pondération lexicale directe, chacun des mots anglais $e_i$ est généré par des mots étrangers alignés $f_j$ avec la probabilité de traduction lexicale $w(e_i|f_j)$. Et de même pour la pondération lexicale inverse $lex(\bar{f}|\bar{e}, a)$. Les mêmes traits ont été calculés pour les mots pleins lemmatisés. Ce trait pourrait refléter la confiance de l'alignement entre les deux segments.
11) La somme de différence de probabilités de traduction lexicale entre la traduction humaine et la traduction la plus probable selon la table de probabilité. Pour chaque mot source, nous prenons le mot cible dans la traduction humaine avec la plus grande probabilité. Par exemple pour la paire *alternatives* → *solutions de remplacement*, la traduction la plus littérale est *alternatives* avec une probabilité de 0,4. Dans la traduction humaine, le mot *solutions* possède la plus grande probabilité, mais qui est seulement 0,07. Selon cette méthode, nous comptons aussi les mots non alignés pour calculer un ratio sur le nombre total de tokens de chaque côté. Ces traits ont été calculés dans les deux directions de traduction.

Le nombre d'instances pour la validation croisée est assez limité, nos expériences utilisant des classifieurs d'apprentissage statistique obtiennent de meilleurs résultats qu'avec les réseaux neuronaux (Zhai *et al.*, 2019). La boîte à outils *Scikit-Learn* (Pedregosa *et al.*, 2011) est utilisée pour entraîner différents classifieurs [13].

8. La lemmatisation anglaise est faite par *Stanford CoreNLP*, celle française par *Tree Tagger* (Schmid, 1995), puisque ce n'est pas encore possible par *Stanford CoreNLP*.
9. `https://github.com/commonsense/conceptnet5/wiki/Downloads`
10. Nous filtrons les mots selon une liste manuelle, qui contient les verbes légers, déterminants, pronoms, etc.
11. `https://wit3.fbk.eu/`, `https://paracrawl.eu/index.html`
12. $lex(\bar{e}|\bar{f}, a) = \prod_{i=1}^{length(\bar{e})} \frac{1}{|\{j|(i,j)\in a\}|} \sum_{\forall(i,j)\in a} w(e_i|f_j)$
13. Le jeu de données et le code sont disponibles ici : `https://github.com/YumingZHAI/ctp`.

# 5 Résultats expérimentaux et analyse

Le nombre d'instances de *non_littéral* (1127) est seulement un tiers de *littéral* (3771). Compte tenu de cet écart, nous avons évalué les classifieurs sous plusieurs configurations : ($a$) six classes (*littéral, équivalence, généralisation, particularisation, modulation, contient_transposition*), où *littéral* contient toutes les instances, ou 200 instances pour une distribution approximativement équilibrée [14]. ($b$) deux classes (*littéral* et *non_littéral*), avec trois répartitions (3 :1, 2 :1, 1 :1) ($c$) cinq classes : seulement les catégories non littérales.

Ces classifieurs ont été entraînés : *RandomForest, Multilayer Perceptron, Logistic Regression, Support Vector Machine, K-nearest Neighbors, Decision Tree, Bernoulli Naive Bayes, multinomial Naive Bayes* et *Gaussian Naive Bayes*. Pour chaque configuration, nous avons optimisé les hyperparamètres de ces classifieurs [15]. L'évaluation est menée par une validation croisée à cinq plis (en utilisant *StratifiedKFold*), selon les mesures de l'exactitude (*accuracy*) moyenne, la F-mesure micro-moyenne et macro-moyenne (Tsoumakas *et al.*, 2011). Les résultats sous différentes configurations sont récapitulés dans la table 2, où le classifieur *Dummy* est utilisé comme une baseline, qui prédit toujours la classe la plus nombreuse. Pour toutes les configurations, le classifieur *RandomForest* obtient toujours la meilleure performance [16].

Nous essayons d'abord une classification directe en six classes. Les résultats de notre classifieur dépassent largement ceux du classifieur *Dummy*. En revanche, la difficulté de la tâche en multi-classe est aussi reflétée dans la distribution approximativement équilibrée. Ainsi nous décidons de diviser le problème : effectuer d'abord une classification binaire, suivi par une classification multi-classe parmi les catégories non littérales.

Pour la classification binaire, les deux meilleurs classifieurs sont *RandomForest* et *Multilayer Perceptron*. En plus, *RandomForest* est meilleur que les deux assemblés par la méthode *hard voting* ou *soft voting*. De la distribution naturelle (3 :1) à la distribution équilibrée (1 :1), la F-mesure moyenne pour la classe *non_littéral* augmente de 0,78 à 0,88. Nous continuerons à tester cette tendance quand un jeu de données plus large sera disponible. Une analyse des erreurs sur la distribution 3 :1 montre que parmi les 290 instances *non_littéral* classifiées en *littéral*, 117 sont de classe *équivalence*. Cela indique que ces deux classes sont difficiles à distinguer pour le classifieur.

L'exactitude la plus élevée pour la classification entre les classes non littérales est de 55,10%. Des F-mesure moyennes sur les cinq plis pour chaque classe sont : *équivalence* (0,51), *généralisation* (0,25), *particularisation* (0,56), *modulation* (0,36) et *contient_transposition* (0,68). La catégorie *généralisation* contient beaucoup moins d'instances que les autres catégories, qui nécessite une augmentation ; il existe beaucoup de confusion entre *modulation* et les autres catégories, qui suggère une amélioration du guide d'annotation ; la confusion existe aussi entre *équivalence* et *contient_transposition*.

Avec le meilleur classifieur *RandomForest*, nous avons effectué une étude d'ablation de traits. Pour la classification binaire, le groupe de traits *alignement de mot* contribue le plus. Pour la classification en cinq classes, la combinaison de tous les traits sauf le groupe *ressource externe* génère le meilleur résultat (exactitude moyenne 55,20%), où les groupes *analyse morpho-syntaxique* et *analyse syntaxique* contribuent plus. En général, des traits en nombre réel ont des meilleures performances que

14. Des instances de *littéral* ont été extraites au hasard pour les configurations $a$ et $b$.

15. Pour trouver les meilleurs hyperparamètres, 10% de données sont séparées comme test, et une validation croisée à trois plis est exécutée sur 90% de données d'entraînement.

16. Les hyperparamètres en détail sont donnés ensemble avec le code.

des traits en nombre entier.

| Distribution de classes | Classifieur | Exactitude moyenne | Micro-F1 | Macro-F1 |
|---|---|---|---|---|
| **Six classes** | | | | |
| six classes, avec 3771 *littéral* | Dummy | 76,99% | 0,77 | 0,14 |
| | RandomForest | **83,10%** | 0,83 | 0,44 |
| six classes, avec 200 *littéral* | Dummy | 25,77% | 0,26 | 0,07 |
| | RandomForest | 57,04% | 0,57 | 0,52 |
| **Deux classes** | | | | |
| *littéral* (3) : *non_littéral* (1) | Dummy | 76,99% | 0,77 | 0,43 |
| | RandomForest | **90,16%** | 0,90 | 0,86 |
| *littéral* (2) : *non_littéral* (1) | Dummy | 66,67% | 0,67 | 0,40 |
| | RandomForest | 88,85% | 0,89 | 0,88 |
| *littéral* (1) : *non_littéral* (1) | Dummy | 50,00% | 0,50 | 0,33 |
| | RandomForest | **87,09%** | 0,87 | 0,87 |
| **Cinq classes** | | | | |
| Cinq classes non_littéral | Dummy | 30,35% | 0,30 | 0,09 |
| | RandomForest | **55,10%** | 0,55 | 0,47 |

TABLE 2 – Résultats expérimentaux sous différentes configurations, utilisant tous les traits

# 6 Conclusion et perspectives

En nous fondant sur notre corpus annoté manuellement en procédés de traduction au niveau sous-phrastique, nous avons proposé une classification automatique. Avec les traits implémentés et par le classifieur *RandomForest*, l'exactitude la plus élevée est de 87,09% pour la classification binaire (distribution équilibrée), et de 55,20% pour la classification entre les procédés de traduction non littérale. Les résultats de notre classifieur sont encourageants et nous continuerons à l'améliorer. Nous utiliserons cette connaissance pour mieux contrôler le processus d'extraction de paraphrases à partir de corpus parallèle bilingues. Il est aussi pertinent de l'intégrer dans le processus de traduction automatique pour mieux traiter les traductions non littérales, ou de l'utiliser pour assister aux études en traductologie.

La classe *Généralisation* contient beaucoup moins d'instances que les autres classes. Nous aurons recours à la ressource PPDB, ConceptNet et Linguee pour construire un jeu de données plus équilibré. L'annotation manuelle se poursuivra pour fournir plus de données, surtout pour le couple anglais-chinois. Nous exploiterons d'autres traits pour mieux effectuer la classification multi-classe, tels que la probabilité de traduction des segments, la liste des expressions figées, etc. Une perspective importante est d'étendre ce travail sur des corpus parallèles non alignés manuellement au préalable. Cette configuration demandera un alignement de mot automatique de bonne performance : d'abord aligner les traductions littérales mot à mot, ensuite aligner des blocs *n-m* sur des couples de segments traduits non littéralement.

# Remerciements

Nous remercions les relecteurs anonymes pour leurs nombreuses remarques constructives. Nous exprimons aussi notre gratitude à Aurélien Max pour ses propositions des traits pertinents à implémenter, et à Cyril Grouin pour ses conseils en vue d'améliorer le guide d'annotation.

# Références

BANNARD C. & CALLISON-BURCH C. (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, p. 597–604 : Association for Computational Linguistics.

CAMACHO-COLLADOS J., PILEHVAR M. T., COLLIER N. & NAVIGLI R. (2017). Semeval-2017 task 2 : Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, p. 15–26 : Association for Computational Linguistics.

CANDITO M., NIVRE J., DENIS P. & ANGUIANO E. H. (2010). Benchmarking of statistical dependency parsers for French. In *Proceedings of the 23rd International Conference on Computational Linguistics : Posters*, p. 108–116 : Association for Computational Linguistics Chinese Information Processing Society of China.

CARL M. & SCHAEFFER M. J. (2017). Why Translation Is Difficult : A Corpus-Based Study of Non-Literality in Post-Editing and From-Scratch Translation. *HERMES-Journal of Language and Communication in Business*, (56), 43–57.

CARPUAT M., VYAS Y. & NIU X. (2017). Detecting cross-lingual semantic divergence for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, p. 69–79 : Association for Computational Linguistics.

CHUQUET H. & PAILLARD M. (1989). *Approche linguistique des problèmes de traduction anglais-français*. Ophrys.

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37–46.

DENG D. & XUE N. (2017). Translation Divergences in Chinese-English Machine Translation : An Empirical Investigation. *Computational Linguistics*, **43**(3), 521–565.

DORR B. J., PEARL L., HWA R. & HABASH N. (2002). Duster : A method for unraveling cross-language divergences for statistical word-level alignment. In *Conference of the Association for Machine Translation in the Americas*, p. 31–43 : Springer.

GANITKEVITCH J., VAN DURME B. & CALLISON-BURCH C. (2013). PPDB : The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 758–764.

GRAY R. M. (1990). *Entropy and Information Theory*. Berlin, Heidelberg : Springer-Verlag.

KOEHN P., OCH F. J. & MARCU D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, p. 48–54 : Association for Computational Linguistics.

LEVENSHTEIN V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, **10**(8), 707–710.

LIANG P., TASKAR B. & KLEIN D. (2006). Alignment by agreement. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, p. 104–111 : Association for Computational Linguistics.

MANNING C. D., SURDEANU M., BAUER J., FINKEL J., BETHARD S. J. & MCCLOSKY D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, p. 55–60.

MOLINA L. & HURTADO ALBIR A. (2002). Translation Techniques Revisited : A Dynamic and Functionalist Approach. *Meta*, **47**(4), 498–512.

NEWMARK P. (1981). *Approaches to Translation (Language Teaching Methodology Senes)*. Oxford : Pergamon Press.

NEWMARK P. (1988). *A textbook of translation*, volume 66. Prentice Hall New York.

PAVLICK E., BOS J., NISSIM M., BELLER C., VAN DURME B. & CALLISON-BURCH C. (2015a). Adding semantics to data-driven paraphrasing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, volume 1, p. 1512–1522.

PAVLICK E., RASTOGI P., GANITKEVITCH J., DURME B. V. & CALLISON-BURCH C. (2015b). PPDB 2.0 : Better Paraphrase Ranking, Fine-Grained Entailment Relations, Word Embeddings, and Style Classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2 : Short Papers*, p. 425–430.

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.

PETROV S., DAS D. & MCDONALD R. T. (2012). A Universal Part-of-Speech Tagset. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, M. U. DOGAN, B. MAEGAARD, J. MARIANI, J. ODIJK & S. PIPERIDIS, Eds., *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, p. 2089–2096 : European Language Resources Association (ELRA).

PHAM M. Q., CREGO J., SENELLART J. & YVON F. (2018). Fixing Translation Divergences in Parallel Corpora for Neural MT. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 2967–2973.

SCHMID H. (1995). Improvements In Part-of-Speech Tagging With an Application To German. In *Proceedings of the ACL SIGDAT-Workshop*, p. 47–50.

SPEER R., CHIN J. & HAVASI C. (2017). Conceptnet 5.5 : An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*, p. 4444–4451.

SPEER R. & LOWRY-DUDA J. (2017). Conceptnet at semeval-2017 task 2 : Extending word embeddings with multilingual relational knowledge. In S. BETHARD, M. CARPUAT, M. APIDIANAKI, S. M. MOHAMMAD, D. M. CER & D. JURGENS, Eds., *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*, p. 85–89 : Association for Computational Linguistics.

TSOUMAKAS G., KATAKIS I. & VLAHAVAS I. (2011). Random k-labelsets for multilabel classification. *IEEE Transactions on Knowledge and Data Engineering*, **23**(7), 1079–1089.

VINAY J.-P. & DARBELNET J. (1958). *Stylistique comparée du français et de l'anglais : méthode de traduction*. Bibliothèque de stylistique comparée. Didier.

VYAS Y., NIU X. & CARPUAT M. (2018). Identifying Semantic Divergences in Parallel Text without Annotations. In M. A. WALKER, H. JI & A. STENT, Eds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, p. 1503–1515 : Association for Computational Linguistics.

ZHAI Y. (2018). Construction d'un corpus multilingue annoté en relations de traduction. In *Rencontre Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, p. 85–99, Rennes, France.

ZHAI Y., MAX A. & VILNAT A. (2018). Construction of a Multilingual Corpus Annotated with Translation Relations. In *First Workshop on Linguistic Resources for Natural Language Processing*, p. 102–111, Santa Fe, New Mexico, USA.

ZHAI Y., SAFARI P., ILLOUZ G., ALLAUZEN A. & VILNAT A. (2019). Towards Recognizing Phrase Translation Processes : Experiments on English-French. *CoRR*, **abs/1904.12213**.

# Combien d'exemples de tests sont-ils nécessaires à une évaluation fiable ? Quelques observations sur l'évaluation de l'analyse morpho-syntaxique du français.

Guillaume Wisniewski
LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91 405 Orsay, France
guillaume.wisniewski@limsi.fr

RÉSUMÉ

L'objectif de ce travail est de présenter plusieurs observations, sur l'évaluation des analyseurs morpho-syntaxique en français, visant à remettre en cause le cadre habituel de l'apprentissage statistique dans lequel les ensembles de test et d'apprentissage sont fixés arbitrairement et indépendemment du modèle considéré. Nous montrons qu'il est possible de considérer des ensembles de test plus petits que ceux généralement utilisés sans conséquences sur la qualité de l'évaluation. Les exemples ainsi « économisés » peuvent être utilisés en apprentissage pour améliorer les performances des systèmes notamment dans des tâches d'adaptation au domaine.

ABSTRACT

**Some observations on the evaluation of PoS taggers**

This work aims at reporting several observations on the evaluation of PoS taggers that are challenging the usual framework of statistical learning in which the test sets and are fixed arbitrarily and independently of the model considered. We show that, in many cases, it is possible to consider smaller test sets than those usually used with no impact on the quality of the evaluation.

MOTS-CLÉS : Apprentissage statistique, évaluation.

KEYWORDS: Machine Learning, Evaluation.

## 1 Introduction

L'apprentissage statistique est devenu la solution de choix pour la majorité des problèmes de traitement automatique des langues (TAL) : depuis que Charniak a montré que, pour l'analyse syntaxique, une méthode statistique surpassait les approches à base de règles (Charniak, 1996), on ne compte plus les tâches (identification de la polarité, reconnaissance d'entités nommées, traduction automatique, ...) pour lesquelles les meilleures performances (aussi bien évaluées automatiquement en comparant les sorties des systèmes à des références que par une évaluation qualitative réalisée à la main) sont obtenues en estimant les paramètres d'un modèle sur un corpus d'apprentissage.

La quasi totalité des travaux publiés aujourd'hui considère un même cadre expérimental pour évaluer aussi bien les modèles que les idées proposées : il existe, pour la plupart des tâches, des corpus de « références » (par exemple, pour l'analyse syntaxique, le Penn Tree Bank ou, pour la traduction automatique, les corpus des campagnes d'évaluation WMT) qui définissent un jeu de données d'apprentissage sur lesquels les modèles sont appris et un jeu de test réservé à l'évaluation des

performances de ceux-ci [1].

La répartition des données entre le corpus de test et le corpus d'apprentissage est généralement complètement arbitraire : des décisions aussi importantes que le nombre d'exemples que doivent contenir chacun de ces corpus reposent souvent sur des règles ou des savoir-faire empiriques voire complètement arbitraires [2] et ne sont que très rarement explicitées. Avoir des corpus de test et d'apprentissage fixes et clairement identifiés est supposé garantir facilement que les résultats publiés dans différents articles sont directement comparables et le jeu auquel nous jouons dans la plupart de nos publications consiste à améliorer les résultats de méthodes de références (les fameuses *baselines*) sur un ensemble de test donné et immuable.

Ce cadre expérimental ne correspond cependant pas à la plupart des applications « réelles » des méthodes de TAL : dans de nombreux cas, pour des raisons de coût ou de compétences, seules quelques dizaines voire quelques centaines d'exemples peuvent être étiquetées. Ceux-ci sont généralement réservés à l'évaluation des performances des modèles : constituer un ensemble d'apprentissage de taille suffisante (généralement plusieurs milliers voire dizaines de milliers d'exemples) est tout simplement irréaliste. C'est par exemple le cas lorsque l'on cherche à développer des modèles de TAL pour des langues peu dotées, notamment pour développer des outils d'aide à leur documentation (Michaud *et al.*, 2018), ou à des domaines pour lesquels il n'existe pas de données annotées (Sokolov *et al.*, 2017).

L'objectif de ce travail est de présenter plusieurs observations, sur une application particulière du TAL, l'analyse morpho-syntaxique du français, visant à remettre en cause le cadre habituel de l'apprentissage statistique dans lequel les ensembles de test et d'apprentissage sont fixés arbitrairement et indépendemment du modèle considéré. Les résultats présentés peuvent toutefois se généraliser facilement à tout système de TAL pouvant être évalué par un coût 0/1, Nous montrons que, bien souvent, il est possible de considérer des ensembles de test plus petits que ceux généralement définis sans conséquence sur la qualité de l'évaluation et que les exemples ainsi « économisés » peuvent être utilisés en apprentissage pour améliorer les performances des systèmes notamment dans des tâches d'adaptation au domaine.

# 2 Contexte

## 2.1 Cadre expérimental

**Données** Toutes les expériences présentées dans ce travail ont été réalisées avec les données du projet *Universal Dependencies* [3] (Nivre *et al.*, 2017). Ce projet a pour objectif de développer des corpus étiquetés avec des informations morpho-syntaxiques pour un large éventail de langues. La dernière version de l'UD rassemble 133 corpus couvrant 75 langues. Le projet contient 7 corpus pour le français [4]. Ces corpus présentent une très grande variabilité dans les tailles des différents jeu de données utilisés : les corpus de test comportent entre 110 phrases (2 824 mots) et 2 541 phrases

1. Par soucis de clarté, nous ne parlerons pas, dans ce travail, des corpus de développement et supposerons qu'ils sont intégrés aux corpus d'apprentissage.

2. comme la règle bien connue : « 80% des données pour l'apprentissage et 20% pour le test ».

3. Nous avons utilisé la version 2.3 des données.

4. Le projet contient notamment les conversions du *French Treebank* (Abeillé *et al.*, 2003) et du corpus Sequoia (Candito & Seddah, 2012) dans le formalisme UD ainsi que des corpus collectés spécifiquement comme ParTuT développé à l'université de Turin.

| | FTB | GSD | PUD | ParTUT | SRCMF | Sequoia | Spoken |
|---|---|---|---|---|---|---|---|
| Sequoia | 92,4 $\pm$ 0,1 | 92,3 $\pm$ 0,5 | 86,3 $\pm$ 0,4 | 91,0 $\pm$ 1,1 | 52,3 $\pm$ 0,7 | 96,0 $\pm$ 0,3 | 80,5 $\pm$ 0,7 |
| ParTUT | 88,9 $\pm$ 0,2 | 89,2 $\pm$ 0,6 | 84,3 $\pm$ 0,4 | 94,2 $\pm$ 0,9 | 42,9 $\pm$ 0,7 | 88,6 $\pm$ 0,6 | 76,3 $\pm$ 0,8 |
| GSD | 93,2 $\pm$ 0,1 | 96,2 $\pm$ 0,3 | 89,7 $\pm$ 0,3 | 93,0 $\pm$ 1,0 | 54,4 $\pm$ 0,7 | 94,6 $\pm$ 0,4 | 83,8 $\pm$ 0,7 |
| FTB | 97,1 $\pm$ 0,1 | 93,1 $\pm$ 0,5 | 87,1 $\pm$ 0,4 | 93,5 $\pm$ 0,9 | 54,9 $\pm$ 0,7 | 94,4 $\pm$ 0,4 | 81,2 $\pm$ 0,7 |
| Spoken | 67,9 $\pm$ 0,3 | 69,4 $\pm$ 0,9 | 70,0 $\pm$ 0,5 | 74,6 $\pm$ 1,6 | 48,4 $\pm$ 0,7 | 70,4 $\pm$ 0,9 | 92,3 $\pm$ 0,5 |
| SRCMF | 60,8 $\pm$ 0,3 | 61,5 $\pm$ 0,9 | 63,4 $\pm$ 0,6 | 63,5 $\pm$ 1,8 | 92,5 $\pm$ 0,3 | 62,1 $\pm$ 0,9 | 65,1 $\pm$ 0,9 |

TABLE 1 – Précision (en %) d'un analyseur syntaxique appris et évalué sur les différentes combinaison d'ensemble de test et ensemble d'apprentissage des corpus français du projet Universal Dependencies. Les intervalles de confiance sont calculés en appliquant la méthode de Clopper-Pearson (c.f. § 2.2).

(82 440 mots), les corpus d'apprentissage entre 803 phrases (25 729 mots) et 14 759 phrases (485 464 mots). La table 1 rapporte les performances obtenues par un étiqueteur morpho-syntaxique utilisant un modèle à base d'historique (décrit dans le paragraphe suivant). Elle montre que, conformément à notre intuition, les performances chutent dès que l'on change de domaine et, surtout, que la confiance que l'on a dans l'estimation de la précision varie fortement selon les corpus.

**Analyseur morpho-syntaxique** Nos expériences utilisent un analyseur morpho-syntaxique à base d'historique (Black *et al.*, 1992). Dans ces modèles, la prédiction d'une séquence d'étiquettes morpho-syntaxiques se réduit à une succession de problèmes de classification multi-classe : les étiquettes des mots de la phrase sont prédites l'une après l'autre par un perceptron moyenné. Nous utilisons un jeu de caractéristiques standard (Zhang & Nivre, 2011). Une description détaillée de cet analyseur est faite dans (Bartenlian *et al.*, 2017; Wisniewski *et al.*, 2014b,a). Ce modèle permet d'atteindre des performances proches de l'état de l'art tout en étant extrêmement rapide à entraîner, ce qui permet de multiplier les expériences : il obtient une précision moyenne de 91,10% sur l'ensemble des corpus du projet UD (Aufrant *et al.*, 2017), un résultat comparable au 92.22% obtenu par l'analyseur UDPIPE (Straka & Straková, 2017) utilisé comme système de référence dans le défi *Multilingual Parsing from Raw Text to Universal Dependencies* organisé dans le cadre de CoNLL'17.

Nous utilisons également comme point de comparaison l'analyseur morpho-syntaxique du projet `CoreNLP` développé à Stanford. Cet analyseur repose sur un modèle à maximum d'entropie et utilise un ensemble de caractéristiques riches et des dépendances cycliques (Toutanvoa & Manning, 2000; Toutanova *et al.*, 2003) dont les paramètres ont été estimés sur un des corpus du projet UD [5].

## 2.2 Évaluation d'un classifieur

Un analyseur morpho-syntaxique peut être vu (c'est, en tout cas, de cette manière qu'il est évalué) comme un classifieur prédisant pour chaque mot en contexte, son étiquette morpho-syntaxique et évalué exactement comme un classifieur multi-classe. Nous rappelons rapidement dans ce paragraphe le principe de l'évaluation d'un tel classifieur et expliquons comment la notion d'*intervalle de confiance* permet de mesurer la qualité de cette évaluation.

La qualité d'un classifieur $f$, c'est-à-dire sa capacité à prédire l'étiquette associée à une observation donnée (un mot dans le cas d'un analyseur morpho-syntaxique) est naturellement évaluée par l'*erreur*

5. La documentation du projet ne précise ni la version ni le corpus qui a été utilisé en apprentissage

*en généralisation* :

$$e_g = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(f(x), y)\right] \quad (1)$$

où $\mathcal{D}$ est la distribution selon laquelle les données sont générées et $\ell$ la fonction de coût du problème. Cette erreur ne peut, bien évidemment, pas être calculée directement puisque ce calcul nécessiterait de connaître l'ensemble des données possibles et leur étiquette. Il est toutefois possible de l'estimer sur un échantillon $(x_i, y_i)_{i=1}^n$ de $n$ exemples étiquetés :

$$\hat{e}_n = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i) \quad (2)$$

Si ces données n'ont pas été utilisées pour choisir les paramètres du classifieur, alors $\hat{e}_n$ est un estimateur non biaisé de l'erreur en généralisation (Duda *et al.*, 2001).

Les données utilisées pour estimer l'erreur en généralisation sont habituellement choisies à priori et indépendamment du classifieur et ce choix est fixe : pour la plupart des tâches existantes, la séparation entre ensemble de test et ensemble d'apprentissage est fixée par les personnes ayant collecté les données ou ayant défini la tâche et ce choix n'est quasiment jamais remis en question.

La valeur de $\hat{e}$ va naturellement dépendre de l'échantillon (c.-à-d. du choix du corpus de test) sur lequel elle est estimée et $\hat{e}$ peut être modélisé par une variable aléatoire distribuée selon une loi binomiale. Il est possible de caractériser la qualité d'une estimation, c'est-à-dire la différence entre la valeur réelle d'un paramètre (dans notre cas, l'erreur en généralisation) et son estimation (ici, l'erreur calculée sur un ensemble de test) en construisant un intervalle de confiance de niveau donné (Wasserman, 2013) qui définit une *marge d'erreur* entre la valeur estimée sur un échantillon et un relevé exhaustif sur la population totale.

Le niveau d'un intervalle de confiance, généralement exprimé sous la forme d'un pourcentage, minore la probabilité de contenir la valeur à estimer. Par exemple, si $C$ est un intervalle de confiance à 95% du taux d'erreur, alors on sait que si on construit $n$ intervalles de confiance de la même manière (par exemple en ré-échantillonnant l'ensemble de test) alors, pour $n$ suffisamment grand, au moins 95% d'entre eux contiendront la « vraie » valeur du paramètre à estimer, c'est-à-dire l'erreur en généralisation.

Il est possible de construire un intervalle de confiance pour une variable aléatoire binomiale à l'aide de la méthode de Clopper-Pearson (Clopper & Pearson, 1934) dont il existe des implémentations pour la plupart des langages.

# 3 Quelques observations expérimentales

## 3.1 Impact de la taille de l'ensemble de test sur l'évaluation

Pour illustrer l'impact de la taille de l'ensemble de test sur la qualité de l'évaluation nous proposons de réaliser l'expérience suivante : les paramètres d'un analyseur morpho-syntaxique sont estimés sur l'ensemble d'apprentissage du corpus GSD ; les performances de cet analyseur sont ensuite estimées sur le corpus d'apprentissage du corpus FTB en considérant des corpus de test contenant un nombre d'exemples croissant. Utiliser un ensemble d'apprentissage (mais d'un corpus différent !) permet de garantir que l'on dispose de suffisamment de données (l'ensemble d'apprentissage contient 5 fois

plus de phrases que l'ensemble de test) pour mesurer l'impact de la taille de l'ensemble de test. Des résultats similaires à ceux que nous allons exposer dans la suite de cette section ont été obtenus sur les autres combinaisons d'ensemble de test et d'ensemble d'apprentissage. Par soucis de clarté nous ne détaillerons que les résultats obtenus sur les deux plus grands corpus du corpus UD.

Afin de caractériser l'influence du taux d'erreur, nous considérons, dans cette expérience, trois analyseurs différents : `gsd-full`, l'analyseur implémentant un modèle à base d'historique (c.f. section 2.1) appris sur la totalité du corpus d'apprentissage, `gsd-small` le même modèle appris sur la moitié du corpus d'apprentissage et`stanford`, l'analyseur de `CoreNLP` utilisant le modèle pré-entrainé fourni par les développeurs de cet outil.

La figure 1 représente l'évolution de la précision en fonction de la taille du corpus de test et surtout l'intervalle de confiance à 95% correspondant. Comme expliqué à la section 2.2, la largeur de l'intervalle de confiance permet de mesurer la qualité de l'estimation réalisée. La figure 2 montre l'évolution de cette largeur en fonction de la taille de l'ensemble de test.

Il apparaît que la largeur de l'intervalle de confiance diminue très vite avec le nombre de données et, en pratique, on pourrait réduire la taille du corpus d'évaluation de moitié sans impacter sensiblement la qualité de l'estimation et donc de l'évaluation. En effet, la largeur de l'intervalle de confiance estimé à partir de 40 000 mots est sensiblement le même que celle estimée à partir de 80 000 mots (correspondant, *grosso modo* à la taille du corpus de test « officiel ») : les intervalles de confiance sont, respectivement, $[93{,}8\%, 94{,}3\%]$ et $[93{,}6\%, 94{,}0\%]$. En outre, considérer encore plus d'exemples (p. ex. 160 000) ne permet de réduire la largeur de l'intervalle de confiance que de $0{,}1$ point.

Cette observation est renforcée par le fait qu'à partir d'un corpus de test comportant 40 000 mots, les intervalles de confiance des différents analyseurs ne se chevauchent plus et que la différence entre leurs performances est donc statistiquement significative (Wasserman, 2013). Ainsi, indépendamment de la largeur réelle de l'intervalle de confiance, les évaluations réalisées sont suffisamment précises pour permettre de répondre de manière convaincante à l'une des principales motivations de l'évaluation d'un modèle de TAL : est-ce qu'un analyseur améliore les résultats de la *baseline* ?

## 3.2 Conséquences pour les problèmes d'adaptation au domaine

Nous avons montré, dans la section précédente, qu'il était possible de réduire la taille du corpus de test tout en continuant d'évaluer la qualité d'un modèle avec suffisamment de précision pour pouvoir comparer deux modèles de manière fiable. Nous souhaitons illustrer dans cette section, les conséquences de cette observation sur la tâche d'adaptation au domaine pour l'analyse morpho-syntaxique.

Le cadre expérimental généralement considéré dans les tâches d'adaptation au domaine est le suivant : on dispose d'un corpus de données étiquetées suffisamment grand pour permettre l'apprentissage d'un modèle et d'un corpus de test contenant des données issues d'un domaine différent et ne permettant que d'évaluer la dégradation des performances liées au changement de domaine. C'est, par exemple, ce cadre qui est mis en œuvre dans les expériences décrites dans la table 1.

Motivés par les résultats présentés dans la section précédente, nous souhaitons évaluer l'hypothèse suivante : il est possible de continuer à évaluer les performances d'un analyseur morpho-syntaxique hors-domaine en considérant un corpus de test plus petit que ceux habituellement considérés et d'utiliser les exemples ainsi « économisés » en apprentissage.

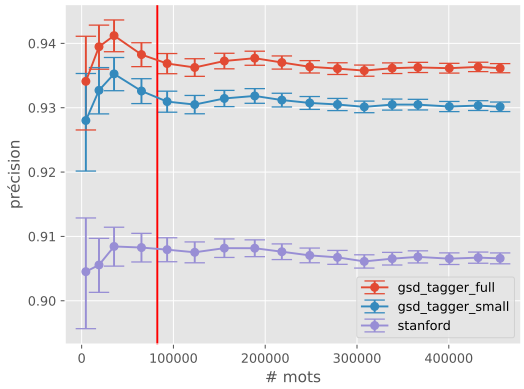



FIGURE 1 – Précision obtenue par 3 analyseurs morpho-syntaxiques estimée sur des corpus de tests de différentes tailles et les intervalles de confiance (à 95%) déterminés par la méthode de Clopper-Pearson correspondant. La ligne rouge correspond à la taille de l'ensemble de test « officiel ».

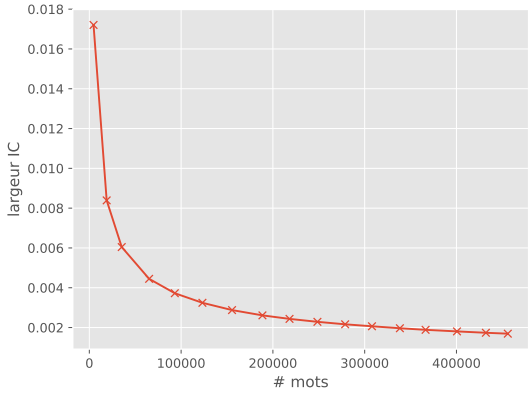



FIGURE 2 – Évolution de la marge d'erreur (largeur de l'intervalle de confiance à 95%) en fonction de la taille du corpus de test.

La figure 3 illustre ce principe. Elle représente l'évolution de l'erreur sur le corpus de test en fonction de la taille de celui-ci pour quatre modèles :

- **GSD** : un modèle à base d'historique appris uniquement sur les données d'apprentissage du corpus GSD (évaluation *out-domain*) ;
- **FTB** un modèle à base d'historique appris uniquement sur les données d'apprentissage sur la totalité du corpus FTB (évaluation *in-domain*) ;
- **FTB-small** un modèle à base d'historique appris uniquement sur les données d'apprentissage issues du corpus FTB et « économisé » en réduisant la taille du corpus d'évaluation ;
- **GSD+FTB-small** un modèle à base d'historique appris sur la concaténation des données GSD et des données économisées en test.

Les résultats présentés à la figure 3 montrent qu'il est possible d'améliorer significativement (les intervalles de confiance ne se chevauchent pas !) les performances d'un analyseurs morpho-syntaxique hors-domaine en utilisant une partie des données étiquetées disponibles pour l'apprentissage plutôt que pour l'évaluation et ce, sans impact sur la confiance que l'on a dans l'estimation de la qualité d'un système.

# 4 Conclusions

Nous avons décrit, dans ce travail, plusieurs observations sur l'évaluation d'un système d'analyse morpho-syntaxiques du français. Ces observations montrent qu'une connaissance, à priori, de l'ordre de grandeur du taux d'erreur permet de réduire de manière significative le nombre d'exemples nécessaires à l'évaluation de l'erreur de généralisation. Même si nous n'avons considéré dans nos expériences que la tâche d'analyse morpho-syntaxique, l'approche décrite peut être appliquée sans problème à toute tâche s'évaluant avec un coût 0/1. Nos travaux futurs porteront, entre autres, sur la possibilité de généraliser notre approche à d'autres fonctions de coût.

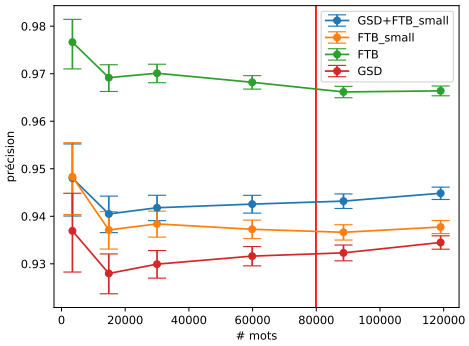



FIGURE 3 – Précision obtenue par un analyseur morpho-syntaxique dans un cadre « adaptation au domaine » en fonction de la taille de l'ensemble d'évaluation.

## Remerciements

Ces travaux ont été en partie financés par l'Agence Nationale de la Recherche (projet PARSITI, ANR-16-CE33-0021). Nous remercions les relecteurs pour leurs commentaires et suggestions.

## Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). *Building a Treebank for French*, In A. ABEILLÉ, Ed., *Treebanks : Building and Using Parsed Corpora*, p. 165–187. Springer Netherlands : Dordrecht.

AUFRANT L., WISNIEWSKI G. & YVON F. (2017). LIMSI@CoNLL'17 : UD shared task. In *Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, p. 163–173, Vancouver, Canada : Association for Computational Linguistics.

BARTENLIAN E., LACOUR M., LABEAU M., ALLAUZEN A., WISNIEWSKI G. & YVON F. (2017). Adaptation au domaine pour l'analyse morpho-syntaxique. In *TALN 2017 - 24e conférence sur le Traitement Automatique des Langues Naturelles*, Orléan, France.

BLACK E., JELINEK F., LAFFERTY J., MAGERMAN D. M., MERCER R. & ROUKOS S. (1992). Towards history-based grammars : Using richer models for probabilistic parsing. In *Proceedings of the Workshop on Speech and Natural Language*, HLT'91, p. 134–139, Stroudsburg, PA, USA : Association for Computational Linguistics.

CANDITO M. & SEDDAH D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *TALN 2012 - 19e conférence sur le Traitement Automatique des Langues Naturelles*, Grenoble, France.

CHARNIAK E. (1996). *Tree-bank Grammars*. Rapport interne, Providence, RI, USA.

CLOPPER C. J. & PEARSON E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, **26**(4), 404–413.

DUDA R. O., HART P. E. & STORK D. G. (2001). *Pattern Classification*. New York : Wiley, 2 edition.

MICHAUD A., ADAMS O., COHN T. A., NEUBIG G. & GUILLAUME S. (2018). Integrating automatic transcription into the language documentation workflow : Experiments with na data and the persephone toolkit. *Language Documentation & Conservation*, p. 393–429.

NIVRE J., AGIĆ Ž., AHRENBERG L. & OTHER (2017). Universal dependencies 2.1. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

SOKOLOV A., KREUTZER J., SUNDERLAND K., DANCHENKO P., SZYMANIAK W., FÜRSTENAU H. & RIEZLER S. (2017). A shared task on bandit learning for machine translation. In *Proceedings of the Second Conference on Machine Translation*, p. 514–524 : Association for Computational Linguistics.

STRAKA M. & STRAKOVÁ J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, p. 88–99, Vancouver, Canada : Association for Computational Linguistics.

TOUTANOVA K., KLEIN D., MANNING C. D. & SINGER Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

TOUTANVOA K. & MANNING C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

WASSERMAN L. (2013). *All of Statistics*. Springer.

WISNIEWSKI G., PÉCHEUX N., GAHBICHE-BRAHAM S. & YVON F. (2014a). Cross-lingual part-of-speech tagging through ambiguous learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1779–1785, Doha, Qatar : Association for Computational Linguistics.

WISNIEWSKI G., PÉCHEUX N., KNYAZEVA E., ALLAUZEN A. & YVON F. (2014b). Apprentissage partiellement supervisé d'un étiqueteur morpho-syntaxique par transfert cross-lingue. In *Proceedings of TALN 2014 (Volume 1 : Long Papers)*, p. 173–183, Marseille, France : Association pour le Traitement Automatique des Langues.

ZHANG Y. & NIVRE J. (2011). Transition-based Dependency Parsing with Rich Non-local Features. In *Proceedings of ACL 2011, the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, p. 188–193, Portland, Oregon, USA : Association for Computational Linguistics.

# De l'extraction des interactions médicament-médicament vers les interactions aliment-médicament à partir de textes biomédicaux: Adaptation de domaine

Tsanta Randriatsitohaina[1] Thierry Hamon[1,2]
(1) LIMSI, CNRS, Université Paris-Saclay, Campus universitaire d'Orsay, 91405 Orsay cedex, France
(2) Université Paris 13, 99 avenue Jean-Baptiste Clément, 93430 Villetaneuse, France
tsanta@limsi.fr, hamon@limsi.fr

## RÉSUMÉ

Les interactions aliments-médicaments (FDI) se produisent lorsque des aliments et des médicaments sont pris simultanément et provoquent un effet inattendu. Nous considérons l'extraction de ces interactions dans les textes comme une tâche d'extraction de relation pouvant être résolue par des méthodes de classification. Toutefois, étant donné que ces interactions sont décrites de manière très fine, nous sommes confrontés au manque de données et au manque d'exemples par type de relation. Pour résoudre ce problème, nous proposons d'appliquer une adaptation de domaine à partir des interactions médicament-médicament (DDI) qui est une tâche similaire, afin d'établir une correspondance entre les types de relations et d'étiqueter les instances FDI selon les types DDI. Notre approche confirme une cohérence entre les 2 domaines et fournit une base pour la spécification des relations et la pré-annotation de nouvelles données. Les performances des modèles de classification appuie également l'efficacité de l'adaptation de domaine sur notre tâche.

## ABSTRACT

**From the extraction of drug-drug interactions to the food-drug interactions in biomedical texts : domain adaptation.**

Food-drug interactions (FDI) occur when food and drug are taken simultaneously and cause unexpected effect. We consider the extraction of these interactions in texts as a relation extraction task that can be resolved with classification methods. Nevertheless, since these interactions are described in a very fine way, we are confronted with a lack of data and a lack of examples per type of relation. To solve this problem, we propose to apply a domain adaptation from drug-drug interactions (DDI) which is a similar task, in order to match relation types and label FDI instances according to DDI types. Our approach confirms consistency between the 2 domains and provides a basis for specifying relations and pre-annotating new data. The performance of classification models also supports the effectiveness of domain adaptation on our task.

MOTS-CLÉS : Adaptation de domaine, Classification, Extraction de relation, Textes biomédicaux.

KEYWORDS: Domain adaptation, Classification, Relation Extraction, Biomedical texts.

# 1 Introduction

Bien qu'il existe des bases ou des terminologies recensant les connaissances d'un domaine de spécialité, disposer d'informations à jour nécessite souvent le recours à la consultation d'articles scientifiques. Ce constat est d'autant plus vrai lorsque les connaissances à recenser ne sont pas déjà présentes dans une base. Ainsi, si les interactions entre médicaments (Aagaard & Hansen, 2013) ou les effets indésirables d'un médicament (Aronson & Ferner, 2005) sont répertoriés dans des bases telles que DrugBank [1] ou Theriaque [2], d'autres informations comme les interactions entre un médicament et un aliment y sont très peu présentes et souvent fragmentées et dispersées dans des sources hétérogènes, principalement sous forme textuelle. Afin de répondre à ces problématiques de mises à jour ou de recensement de ces informations, des méthodes de fouille de textes sont généralement mises en œuvre (Cohen & Hunter, 2008; Rzhetsky *et al.*, 2009; Chowdhury *et al.*, 2011).

Dans cet article, nous nous intéressons à l'identification automatique de mentions d'interaction entre un médicament et un aliment (Food-Drug Interaction - FDI) dans des résumés d'articles scientifiques issus de la base Medline. A l'instar des interactions entre médicaments, une interaction entre un médicament et un aliment correspond à l'apparition d'un effet non attendu quand leur prise est combinée. Par exemple, le pamplemousse est connu pour avoir un effet inhibiteur sur une enzyme impliquée dans le métabolisme de plusieurs médicaments (Hanley *et al.*, 2011). D'autres aliments peuvent avoir des effets sur l'absorption d'un médicament ou sur sa distribution dans l'organisme (Doogue & Polasek, 2013). Pour extraire ces informations des résumés, nous faisons face à plusieurs difficultés : (1) les médicaments et les aliments sont mentionnés de manière très variable dans les résumés. Il peut s'agir des dénominations communes internationales ou des substances actives de médicaments tandis que pour les aliments, il peut aussi être fait mention d'un nutriment, d'un composant particulier ou d'une famille d'aliment ; (2) les interactions sont décrites de manière assez fine dans le corpus annoté à notre disposition ce qui conduit à un nombre d'exemples peu conséquent ; (3) bien que nous disposons de résumés annotés avec des interactions aliment/médicament, l'ensemble des annotations ne couvrent pas de manière homogène les différents types d'interaction et l'ensemble d'apprentissage est bien souvent déséquilibré.

Nous considérons l'extraction de ces interactions comme une tâche d'acquisition de relation étant donné les mentions d'un aliment et d'un médicament. Afin de répondre à ces difficultés, nous proposons d'utiliser une méthode à base d'apprentissage par transfert en tirant partie des données existantes sur les interactions médicaments-médicaments.

Après avoir présenté un état de l'art des méthodes d'acquisition de relations en corpus de spécialité (section 2), nous décrivons à la section 3, le corpus annoté que nous avons utilisé pour mettre au point notre approche afin d'extraire des interactions entre médicament et aliment (section 4). Puis nous présentons et discutons les résultats obtenus (section 6) et nous concluons (section 7).

# 2 Etat de l'art

Différents types d'approches ont été explorés pour extraire les relations à partir des textes biomédicaux. Certaines approches combinent les patrons avec du CRF pour la reconnaissance des symptômes

1. https://www.drugbank.ca/
2. http://www.theriaque.org

dans les textes biomédicaux (Holat *et al.*, 2016). D'autres approches génèrent automatiquement des données lexicales pour le traitement du texte libre dans les documents cliniques en s'appuyant sur un modèle alignement séquentiel multiple pour identifier des contextes similaires (Meng & Morioka, 2015).

Afin d'extraire les interactions médicament-médicament (DDI) (Kolchinsky *et al.*, 2015) se concentrent sur l'identification des phrases pertinentes et des résumés pour l'extraction des preuves pharmacocinétiques. (Ben Abacha *et al.*, 2015) proposent une approche basée sur du SVM combinant : (i) les caractéristiques décrivant les mots dans le contexte des relations à extraire, (ii) les noyaux composites utilisant des arbres de dépendance. L'union et l'intersection des résultats permettent d'obtenir des F1-mesures de 0,5 et 0,39 respectivement. Une méthode en deux étapes basée également sur le classifieur SVM est proposée par (Ben Abacha *et al.*, 2015) pour détecter les DDI potentiels, puis classer les relations parmis les DDI déjà identifiées. Cette seconde approche permet d'obtenir des F1-mesures de 0,53 et 0,40 sur des résumés Medline et 0,83 et 0,68 sur des documents de DrugBank. (Kim *et al.*, 2015) ont construit deux classifieurs pour l'extraction DDI : un classifieur binaire pour extraire les paires de médicaments en interaction et un classifieur de types DDI pour identifier les catégories de l'interaction. (Cejuela *et al.*, 2018) considèrent l'extraction de la relation de localisation des protéines comme une classification binaire. (Liu *et al.*, 2016) proposent une méthode basée sur CNN pour l'extraction des DDI. Dans leur modèle, les mentions de médicaments dans une phrase sont normalisées de la manière suivante : les deux médicaments considérés sont remplacées par `drug1` et `drug2` respectivement dans l'ordre de leur apparition, et tous les autres médicaments sont remplacés par `drug0`. D'autres travaux utilisent un modèle de réseau neuronal récurrent avec plusieurs couches d'attention pour la classification DDI (Yi *et al.*, 2017; Zheng *et al.*, 2017), ou utilisant des récurrences au niveau des mots et des caractères (Kavuluru *et al.*, 2017) produisant une performance de 0,72. Sun *et al.* (2019) propose une méthode hybride combinant un réseau de neurone récurrent et convolutionnel induisant une amélioration de 3%. Le réseau convolutionnel profond de Dewi *et al.* (2017) permet de couvrir de longues phrases qui ont des jeux de données de DDI typique et obtenir une performance de 0,86.

Une méthode simple d'adaptation de domaine consiste à construire un jeu de données étiqueté pour le domaine cible, puis à ajuster un modèle entraîné avec ce dernier (Jiang & Zhai, 2007; Blitzer *et al.*, 2006; Daumé III *et al.*, 2010). Comme notre problème d'extraction d'interaction médicament-aliment est similaire à l'extraction des DDI et les données DDIs sont mieux fournies, nous proposons une approche basée sur l'adaptation de domaine en projetant des modèles entraînés à partir du DDI sur des données FDI afin d'obtenir de nouvelles étiquettes.

# 3 Corpus

Des études ont déjà été menées sur les FDI durant lesquelles l'ensemble de données POMELO a été développé (Hamon *et al.*, 2017). Ce corpus est constitué de 639 résumés d'articles scientifiques du domaine médical (269 824 mots, 5 752 phrases), collectés à partir du portail PubMed [3] avec la requête : `("FOOD DRUG INTERACTIONS"[MH] OR "FOOD DRUG INTERACTIONS*" ) AND ("adverse effects*")`. Les 639 résumés sont annotés selon 9 types d'entités et 21 types de relations avec Brat (Stenetorp *et al.*, 2012) par un étudiant en pharmacie. Les annotations se concentrent sur des informations sur la relation entre aliments, médicaments et pathologies.

3. `https://www.ncbi.nlm.nih.gov/pubmed/`

Étant donné que nous examinons les interactions aliment-médicament dans cet article, nous avons construit notre ensemble de données en tenant compte de tous les couples de *drug* et *food* ou *food-supplement* à partir des données POMELO. L'ensemble de données qui en résulte est composé de 902 phrases étiquetées avec 13 types de relations : *Relation (Rel)*, *Decrease absorption (Dec)*, *No effect on drug (No)*, *Increase absorption (Inc)*, *Negative effect on drug (Neg)*,*Positive effect on drug (Pos)*, *New side effect (New)*, *Without food (Wout)*, *Improve drug effect (Imp)*, *Slow elimination (Sl-e)*, *Slow absorption (Sl-a)*, *Worsen drug effect (Wors)*, *Speed up absorption(Spd)*. Les statistiques de l'ensemble de données sont données dans la table 1.

| Relation | # | Pourcentage | Relation | # | Pourcentage |
|---|---|---|---|---|---|
| unspecified relation | 530 | 58,8% | no effect on drug | 109 | 12,1% |
| decrease absorption | 53 | 5,9% | improve drug effect | 6 | 0,7% |
| positive effect on drug | 21 | 2,3% | without food | 13 | 1,4% |
| negative effect on drug | 88 | 9,8% | speed up absorption | 1 | 0,1% |
| increase absorption | 39 | 4,3% | worsen drug effect | 8 | 0,9% |
| slow elimination | 15 | 1,7% | new side effect | 4 | 0,4% |
| slow absorption | 15 | 1,7% | | | |
| Total | 902 | | | | 100% |

TABLE 1: Statistique des données - Interactions aliment-médicament

# 4 Méthode

Comme les interactions entre aliments-médicaments sont décrites de manière très fine selon de nombreux types de relations, nous sommes confrontés au manque de données et au manque d'exemples par type de relation comme indiqué dans le tableau 1. Par exemple, la relation *speed up absorption* est représenté par un seul exemple, ce qui ne permet pas d'établir une méthode efficace pour extraire automatiquement la relation considérée. Par ailleurs, la tâche d'identification des FDIs présente des similitudes avec la tâche d'extraction des DDIs, où deux médicaments pris ensemble mènent à une modification de leurs effets. Dans cet article, nous étudions la projection des connaissances sur les DDIs afin d'obtenir de nouveaux types correspondant aux FDIs.

## 4.1 Correspondance de types DDI-FDI

Afin de vérifier la cohérence entre deux domaines, nous proposons une approche permettant de trouver une correspondance de type DDI pour chaque type FDI. Pour ce faire, un modèle est entraîné sur les données DDI. Ensuite chaque type de relation est représenté par un ensemble de descripteurs, formant ainsi une instance par type de relation. Puis nous projetons un modèle entraîné avec les données DDI sur cette représentation afin de déterminer à quel type de DDIs correspond la relation. L'ensemble de données de relation est donc un vecteur $D_R$ = [$D_1$, $D_2$, ...,$D_n$] de taille $n$, où $n$ est le nombre de types de relations FDI, $D_i$ est un ensemble de descripteurs représentant la $i^th$ relation. Dans cet article, nous construisons les descripteurs $D_i$ à partir des descripteurs de chaque phrase $P_i$ étiquetée par la relation $R_i$ dans l'ensemble de données initial $D$. Ainsi, le modèle affecte une étiquette de type DDI pour chaque type FDI.

## 4.2 Contraste des étiquettes d'instances

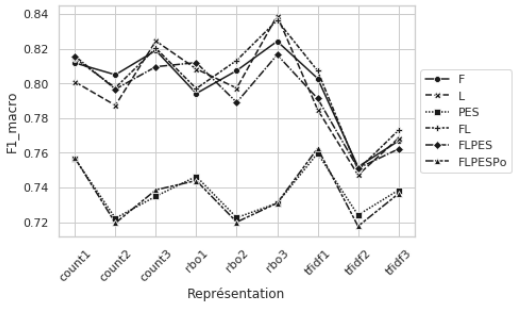
Dans cette section, nous analysons individuellement les étiquettes DDI affectées aux instances FDI afin de déterminer si tous les membres d'une classe FDI donnée sont affectés au même type DDI et appuyer la correspondance obtenu dans la section 4.1. Pour ce faire, nous retenons le meilleur modèle obtenu pour la classification des DDIs et l'appliquons sur les données FDI afin d'obtenir de nouvelles étiquettes pour chaque instance. Cette méthode nous permet de contraster les annotations des données POMELO et une annotation basée sur les types DDIs. Une méthode de classification des instances FDI est ensuite mise en place en utilisant les étiquettes DDI.

# 5 Expérimentation

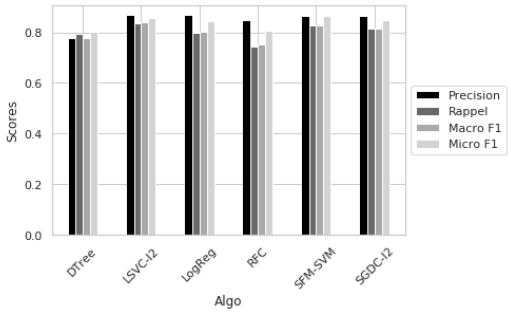
## 5.1 Modèle DDI

L'adaptation de domaine DDI-FDI est évaluée à travers l'utilisation de classifieurs et de descripteurs.

**Corpus** Le corpus DDI [4] est composé de 2280 instances extraites de la base DrugBank et des résumés Medline, similaire au corpus utilisé lors de la compétition (Segura-Bedmar *et al.*, 2013), étiquetées suivant quatre types : *conseil* (540) pour une recommandation concernant l'utilisation concomitante de deux médicaments, *effet* (591) pour l'effet du DDI, *mécanisme* (1006) pour la pharmacodynamique (les effets d'un médicament sont modifiés par la présence d'un autre médicament) ou la pharmacocinétique (les processus par lesquels les médicaments sont absorbés, distribués, métabolisés et excrétés), *interaction* (143) où aucune information sur l'interaction n'est fournie.




(a) Macro F1 obtenu avec SVM selon la représentation des descripteurs.




(b) Performances obtenues selon les modèles, sur l'unigramme, bigramme et trigramme des lemmes.

FIGURE 1: Performances obtenues sur les données DDI.

**Descripteurs.** Dans chaque phrase, les chiffres sont remplacés par le caractère '#' comme proposé dans (Kolchinsky *et al.*, 2015), les autres caractères spéciaux sont supprimés, et chaque mot est converti en minuscule. Nous analysons l'impact de différents descripteurs : formes fléchies (F) des mots, lemmes (L), étiquettes morpho-syntaxiques (Po), fenêtres de mots précédant le premier argument de la relation (P), mots entre les deux arguments (E) et mots suivant le second argument (S).

4. https ://github.com/dbmi-pitt/public-PDDI-analysis/blob/master/PDDI-Datasets/DDI-Corpus2013

**Classification.** La performance de 6 classifieurs est évaluée selon la précision (P), le rappel (R), et la F1-mesure ($F_1$) obtenus par un processus de validation croisée en 10 échantillons. Nous utilisons l'implémentation Scikit-learn (Pedregosa *et al.*, 2011) des classifieurs : (1) un arbre de décision (DTree), (2) un classifieur SVM l2-linéaire (LSVC-l2), (3) une régression logistique (LogReg), (4) un classifieur bayésien naïf multinomial (MNB), (5) une forêt aléatoire (RFC) et (6) un SVM combiné avec un algorithme de sélection de descripteurs (SFM-SVM).

## 5.2 Adaptation de domaine

Afin d'étiqueter les données FDI selon les types DDI, la meilleure configuration obtenue à la section 5.1 est appliquée sur les données POMELO selon la méthode décrite dans la section 4. D'après les nouvelles étiquettes, on obtient une répartition différente des instances : interaction (104 instances), mécanisme (397 instances), conseil (72 instances), effet (329 instances) (table 2). Les modèles de classification sont alors évalués sur nos données en utilisant les nouvelles étiquettes afin de déterminer l'effet de l'adaptation de domaine sur l'identification automatique des relations. De la même manière que pour le modèle DDI décrit à la section 5.1, nous évaluons les 6 classifieurs par un processus de validation croisée en 10 échantillons en faisant varier les descripteurs utilisés.

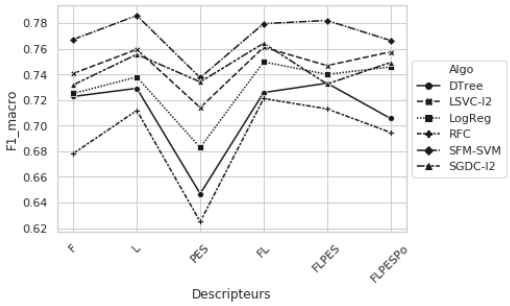
| FDI | Rel | Dec | No | Inc | Neg | Pos | New | Wout | Imp | Sl-e | Wors | Sl-a | Spd |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DDI | M | M | M | M | E | E | E | C | E | M | E | M | M |
| Cons | 7 | 2 | 9 | 5 | 9 | 33 | 0 | **54** | 16 | 0 | 0 | 0 | 0 |
| Effet | **44** | 2 | 19 | 5 | **53** | **38** | **75** | 23 | **67** | 7 | **50** | 0 | 0 |
| Int | 17 | 2 | 1 | 0 | 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Méca | 32 | **94** | **71** | **90** | 22 | 29 | 25 | 23 | 17 | **93** | **50** | **100** | **100** |

TABLE 2: Correspondance type DDI-FDI (Ligne 1) et pourcentage d'instances FDI affectées au type DDI (Ligne 2-5) - *Conseil (C)*, *Mécanisme (M)*, *Effet (E)*, *Interaction (Int)*.

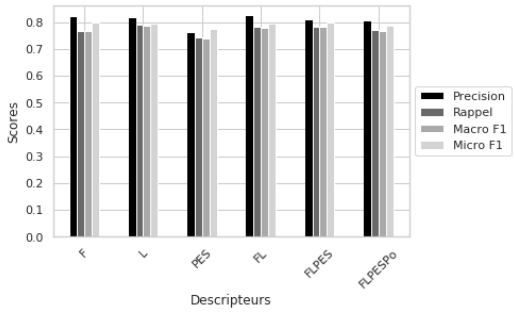
# 6 Résultat

Pour l'extraction des DDIs, nous avons testé l'impact des unigramme, bigramme et trigramme des descripteurs (figure 1a) en les représentant par leur occurence binaire (rbo), leur fréquence (count) ou leur TF-IDF (tfidf). La meilleure F1-mesure 0,839 est obtenue avec un SVM linéaire qui présente un bon rappel (figure 1b) en utilisant les trigrammes des lemmes comme descripteurs. Ce résultat est meilleur que celui obtenu par le meilleur modèle sur la même tâche lors de la compétition Semeval 2013 (Segura-Bedmar *et al.*, 2013). Nous avons appliqué ce modèle sur les représentations des relations FDI comme décrit à la section 4.1 afin d'obtenir une correspondance entre les types DDI et FDI. Les résultats (tableau 2, ligne 1) indiquent une cohérence entre les 2 domaines. En effet, les interactions impliquant la cinétique des médicaments (absorption, élimination, métabolisme) sont étiquetées *Mécanisme*, les interactions impliquant les effets du médicament (positif, négatif, secondaire) sont étiquetées *Effet*, la relation *Without food* qui est une contre-indication de prise d'aliment avec le médicament est étiquetée *Conseil*. Cette cohérence des étiquettes démontrent l'efficacité de l'approche par la représentation des relations. Toutefois, nous remarquons que la *relation* non-spécifiée n'a pas été étiquetée *Int*. L'analyse des instances de cette relation (tableau 2,

lignes 2-5) montre qu'un peu moins de la moitié est considérée comme des effets et le tiers comme des mécanismes. Ces informations pourront servir d'appui pour ajouter des précisions sur les interactions non spécifiées. L'homogénéité élevée pour les relations impliquant des mécanismes suggère une bonne qualité des annotations FDIs pour ces types de relation. Les relations impliquant des effets sont beaucoup moins homogènes, mais les nouvelles étiquettes peuvent servir de base pour une ré-évaluation des annotations des FDIs. Avec les nouvelles étiquettes obtenues, le résultat obtenu sur la tâche d'extraction des FDIs confirme effectivement l'efficacité de la méthode car la F1-mesure résultante est passée de 0,41 sur les étiquettes initiales à 0,79 sur les nouvelles étiquettes (figure 2a). Ce résultat est obtenu par un modèle SVM avec une régularisation l2 précédé d'un processus de sélection de descripteurs SFM, en utilisant les unigramme, bigramme et trigramme des lemmes comme descripteurs. La précision et le rappel sont visiblement équilibrés, ainsi que la macro et micro F1, (figure 2b) ce qui suggère un bon équilibre des données.




(a) Macro F1 selon les différents modèles.




(b) Performances obtenues avec SFM-SVM.

FIGURE 2: Performances obtenues sur les données POMELO.

# 7 Conclusion et Perspectives

Notre article contribue à la tâche d'extraction des interactions aliments-médicaments (FDI) à partir de la littérature scientifique, que nous traitons comme une tâche d'extraction de relation. L'apprentissage supervisé dans ce but n'est pas très concluant car nous sommes confrontés à un problème de manque d'exemples en raison du nombre élevé de types de relations étant donné que les interactions sont décrites de manière très fine. Pour pallier ce problème, nous proposons d'appliquer une adaptation de domaine à partir des interactions médicaments-médicaments (DDI) qui est une tâche similaire, afin d'établir une correspondance entre les types de relations et étiqueter les instances FDI selon les types DDI. Les résultats obtenus suggèrent que l'approche basée sur la représentation des relations à partir des instances est efficace pour identifier les correspondances des types et évaluer la cohérence entre les 2 domaines. Les nouvelles étiquettes obtenues peuvent servir d'appui pour préciser les interactions non spécifiées et pré-annoter de nouvelles données. Pour la suite du travail, nous envisageons d'utiliser d'autre corpus et méthode DDI et confronter les résultats pour une meilleure annotation les FDIs.

# Remerciements

Ce travail est financé par l'ANR dans le cadre du projet MIAM (ANR-16-CE23-0012).

# Références

AAGAARD L. & HANSEN E. (2013). Adverse drug reactions reported by consumers for nervous system medications in europe 2007 to 2011. *BMC Pharmacology & Toxicology*, **14**, 30.

ARONSON J. & FERNER R. (2005). Clarification of terminology in drug safety. *Drug Safety*, **28**(10), 851–70.

BEN ABACHA A., CHOWDHURY M. F. M., KARANASIOU A., MRABET Y., LAVELLI A. & ZWEIGENBAUM P. (2015). Text mining for pharmacovigilance : Using machine learning for drug name recognition and drug-drug interaction extraction and classification. *Journal of Biomedical Informatics*, **58**, 122–132.

BLITZER J., MCDONALD R. & PEREIRA F. (2006). Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, p. 120–128 : Association for Computational Linguistics.

CEJUELA J. M., VINCHURKAR S., GOLDBERG T., PRABHU SHANKAR M. S., BAGHUDANA A., BOJCHEVSKI A., UHLIG C., OFNER A., RAHARJA-LIU P., JENSEN L. J. & ROST B. (2018). LocText : relation extraction of protein localizations to assist database curation. *BMC Bioinformatics*, **19**(1), 15.

CHOWDHURY F. M., LAVELLI A. & MOSCHITTI A. (2011). A study on dependency tree kernels for automatic extraction of protein-protein interaction. In *Proceedings of BioNLP 2011 Workshop*, p. 124–133 : Association for Computational Linguistics.

COHEN K. & HUNTER L. (2008). Getting started in text mining. *PLoS Computational Biology*, **4**(1), e20.

DAUMÉ III H., KUMAR A. & SAHA A. (2010). Frustratingly easy semi-supervised domain adaptation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, p. 53–59 : Association for Computational Linguistics.

DEWI I. N., DONG S. & HU J. (2017). Drug-drug interaction relation extraction with deep convolutional neural networks. *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, p. 1795–1802.

DOOGUE M. & POLASEK T. (2013). The abcd of clinical pharmacokinetics. *Ther Adv Drug Saf*, **4**(1), 5–7.

HAMON T., TABANOU V., MOUGIN F., GRABAR N. & THIESSARD F. (2017). Pomelo : Medline corpus with manually annotated food-drug interactions. In *Proceedings of Biomedical NLP Workshop associated with RANLP 2017*, p. 73–80, Varna, Bulgaria.

HANLEY M., CANCALON P., WIDMER W. & GREENBLATT D. (2011). The effect of grapefruit juice on drug disposition. *Expert Opin Drug Metab Toxicol*, **7**(3), 267–286.

HOLAT P., TOMEH N., CHARNOIS T., BATTISTELLI D., JAULENT M.-C. & MÉTIVIER J.-P. (2016). Weakly-supervised symptom recognition for rare diseases in biomedical text. In *Proceedings of the 15th International Symposium IDA 2016*, Lecture Notes in Computer Science, 9897, Advances in Intelligent Data Analysis XV, p. 192–203, Stockholm, Sweden.

JIANG J. & ZHAI C. (2007). Instance weighting for domain adaptation in nlp. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, p. 264–271 : Association for Computational Linguistics.

KAVULURU R., RIOS A. & TRAN T. (2017). Extracting drug-drug interactions with word and character-level recurrent neural networks. In *Healthcare Informatics (ICHI), 2017 IEEE International Conference on*, p. 5–12 : IEEE.

KIM S., LIU H., YEGANOVA L. & WILBUR W. J. (2015). Extracting drug–drug interactions from literature using a rich feature-based linear kernel approach. *Journal of biomedical informatics*, **55**, 23–30.

KOLCHINSKY A., LOURENÇO A., WU H.-Y., LI L. & ROCHA L. M. (2015). Extraction of pharmacokinetic evidence of drug–drug interactions from the literature. *PloS one*, **10**(5), e0122199.

LIU S., TANG B., CHEN Q. & WANG X. (2016). Drug-drug interaction extraction via convolutional neural networks. *Computational and Mathematical Methods in Medicine*, **2016**.

MENG F. & MORIOKA C. (2015). Automating the generation of lexical patterns for processing free text in clinical documents. *Journal of the American Medical Informatics Association*, **22**(5), 980–986.

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.

RZHETSKY A., SERINGHAUS M. & GERSTEIN M. B. (2009). Getting started in text mining : Part two. *PLoS Comput Biol*, **5**(7), e1000411.

SEGURA-BEDMAR I., MARTÍNEZ P. & HERRERO ZAZO M. (2013). Semeval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2 : Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, p. 341–350 : Association for Computational Linguistics.

STENETORP P., PYYSALO S., TOPIĆ G., OHTA T., ANANIADOU S. & TSUJII J. (2012). Brat : A web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, p. 102–107, Stroudsburg, PA, USA : Association for Computational Linguistics.

SUN X., DONG K., MA L., SUTCLIFFE R. F. E., HE F., CHEN S.-S. & FENG J. (2019). Drug-drug interaction extraction via recurrent hybrid convolutional neural networks with an improved focal loss. *Entropy*, **21**, 37.

YI Z., LI S., YU J., TAN Y., WU Q., YUAN H. & WANG T. (2017). Drug-drug interaction extraction via recurrent neural network with multiple attention layers. In *International Conference on Advanced Data Mining and Applications*, p. 554–566 : Springer.

ZHENG W., LIN H., LUO L., ZHAO Z., LI Z., ZHANG Y., YANG Z. & WANG J. (2017). An attention-based effective neural model for drug-drug interactions extraction. In *BMC Bioinformatics*.

# Demonette2 - Une base de données dérivationnelles du français à grande échelle : premiers résultats

Fiammetta Namer[1] Lucie Barque[2] Olivier Bonami[2] Pauline Haas[3] Nabil Hathout[4]
Delphine Tribout[5]
(1) UMR 7118 ATILF, Nancy, (2) UMR 7110 LLF, Paris
(3) UMR 8094 Lattice, Paris, (4) UMR 5263 CLLE-ERSS, Toulouse
(5) UMR 8163 STL, Lille
fiammetta.namer@univ-lorraine.fr, {lucie.barque;pauline.haas}@univ-paris13.fr, olivier.bonami@linguist.univ-paris-diderot.fr, Nabil.Hathout@univ-tlse2.fr, delphine.tribout@univ-lille3.fr

RÉSUMÉ

Cet article présente la conception et le développement de Demonette2, une base de données dérivationnelle à grande échelle du français, développée dans le cadre du projet ANR Démonext (ANR-17-CE23-0005). L'article décrit les objectifs du projet, la structure de la base et expose les premiers résultats du projet, en mettant l'accent sur un enjeu crucial : la question du codage sémantique des entrées et des relations.

ABSTRACT

**Demonette2 – A large scale derivational database for French: first results**

This paper presents the design and development of Demonette2, a large-scale derivational database of French, developed as part of the ANR Démonext project (ANR-17-CE23-0005). It describes the objectives of the project, the structure of the database and presents the first results of the project, focusing on the question of the semantic encoding of lexical units and their relationships.

MOTS-CLES : base de données dérivationnelles, codage sémantique, paradigmes dérivationnels, lexique français.

KEYWORDS: derivational database, semantic encoding, derivational paradigms, French lexicon.

# 1 Introduction

Démonette2 est une base de données morphologiques (BDM) qui décrit les propriétés dérivationnelles des mots du français de manière extensive et systématique. Alimentée par des ressources lexicales existantes de nature variées, cette base constitue une combinaison inédite d'informations répondant à des besoins multiples, comme la confirmation empirique et l'élaboration d'hypothèses en morphologie, le développement d'outils en traitement automatique des langues, l'enseignement du vocabulaire et le traitement des troubles du langage développementaux ou acquis. Développée dans le cadre du projet Démonext[1] (2018-2021), cette base décrira à terme un réseau dérivationnel totalisant au moins 366 000 entrées, comportant des relations morphologiques directes, indirectes, ascendantes et descendantes et une représentation du sens construit ; les lexèmes seront munis d'annotations morphologiques, de caractérisations sémantiques, de représentations phonologiques, de fréquences d'emploi dans différents corpus,

[1] Démonext bénéficie du soutien de l'ANR 17-CE23-0005, et réunit 4 UMR : ATILF, STL, CLLE-ERSS et LLF.

d'indications de l'âge d'acquisition, etc. Le résultat sera accessible grâce une plateforme offrant un accès adapté à différents publics. La BDM sera distribuée sous licence libre via l'EQUIPEX Ortolang (https://www.ortolang.fr/) et la plateforme REDAC (http://redac.univ-tlse2.fr/).

# 2 Etat de l'art

L'analyse morphologique constitue l'une des étapes initiales centrales des systèmes de TAL. Les analyseurs, le plus souvent basés sur des méthodes d'apprentissage automatique, réalisent un découpage morphématique des mots et permettent de compenser les limitations des lexiques. On citera Linguistica (Goldsmith, 2001), Morfessor (Creutz, 2003 ; Creutz, Lagus, 2005), l'analyseur de Bernhard (2009) ou plus récemment les modèles de Cotterell *et al*. (2015, 2017). Applicables à n'importe quelle langue, ces systèmes sont plus efficaces pour les langues à morphologie concaténative comme l'anglais, l'allemand et le français. Parallèlement, des analyseurs symboliques ont été développés par des linguistes ; pour un panorama, voir Bernhard *et al*. (2011), ou Namer (2013). Les analyseurs morphologiques peuvent être complétés par des ressources lexicales munies d'annotations dérivationnelles, ayant une couverture lexicale assez importante et un ensemble suffisamment riche et varié de propriétés codées pour être exploitables dans une chaîne de traitement en TAL. Malgré un besoin important pour des ressources de ce type, très peu ont été développées au cours des vingt dernières années – et pratiquement aucune pour les langues romanes, comme le constataient déjà Dal *et al*. (1999). La plus connues de ces ressources est CELEX (Baayen *et al*., 1995), qui décrit pour l'allemand, l'anglais et le néerlandais, les propriétés phonétiques, flexionnelles, morpho-syntaxiques, dérivationnelles et statistiques d'un peu moins de 250 000 mots non fléchis issus de dictionnaires et de corpus littéraires et journalistiques. Sinon, citons CatVar (Habash, Dorr, 2003) qui est un système lexical de 100 000 lexèmes de l'anglais réunis en sous-familles dérivationnelles organisées en graphes ; sur le même principe, DerivBase (Zeller *et al*., 2013) contient 215 000 unités lexicales de l'allemand dont les regroupements en familles dérivationnelles sont motivés sémantiquement ; la version 3.0 de WordNet (Fellbaum *et al*., 2007) est enrichie de relations dérivationnelles annotées sémantiquement entre les verbes et une partie de leurs dérivés nominaux (par exemple, la relation $\text{EMPLOY}_{\text{V}}$/$\text{EMPLOYER}_{\text{N}}$ est étiquetée agent). Pour le français, on peut citer deux initiatives récentes visant le développement de réseaux lexicaux à large couverture. Le RL-fr (Lux-Pogadalla, Polguère, 2014) décrit 1 million d'entrées au moyen de relations sémantiques inspirées de (Mel'čuk, 1996). La base JeuxDeMots (Lafourcade, Joubert, 20008), antérieure à RL-fr, est fondée aussi sur le même principe, mais comporte également des relations dérivationnelles. JeuxDeMots obéit à une conception participative de l'enrichissement du réseau, sous forme de jeu en ligne. La fiabilité des termes proposés par les joueurs s'accroît en fonction du nombre de réponses identiques. En 10 ans, la ressource a atteint 270 millions de relations instanciant 151 fonctions lexicales différentes et connectant quelque 3,5 millions de termes. Comme nous le montrons dans la suite, JeuxDeMots et Démonette2 sont en quelque sorte complémentaires puisqu'elles se distinguent en termes de couverture et de mode de développement, dans la mesure où l'évolution de la première dépend de l'imagination des participants, là où la seconde s'appuie sur l'expertise des auteurs des sources d'alimentation de la base pour garantir cohérence et validité théorique à la description morphologique de chaque relation.

La carence de ressources purement dérivationnelles pour le français a motivé le développement, à partir de 2014, du prototype Démonette1 (Hathout, Namer, 2014a, 2014b, 2015, 2016 ; Namer *et al*., 2017). Démonette1 décrit une partie des familles dérivationnelles des verbes, accompagnés de leurs noms d'agent, d'activité et adjectifs de propriété modalisée. Trois objectifs étaient visés : (1)

produire une ressource dont les entrées sont des relations dérivationnelles munies d'annotations riches, notamment sémantiques (Namer, 2002) ; (2) compléter les dérivations base → dérivé par toutes relations motivées qui existent entre membres des familles dérivationnelles, suivant le modèle analogique implémenté dans Morphonette (Hathout, 2009) ; (3) définir une architecture extensible et redondante, qui peut être alimentée par des ressources morphologiques hétérogènes.

Forte de l'expérience acquise avec Démonette1, la BDM Démonette2 construite dans le cadre du projet ANR Démonext se veut à terme une ressource de grande ampleur disposant de descriptions riches des lexèmes, des relations dérivationnelles entre lexèmes et des paradigmes où celles-ci s'insèrent. Démonette2 est par ailleurs compatible avec les principales théories morphologiques actuelles (qu'elles soient morphématiques, lexématiques ou paradigmatiques, cf. respectivement Halle, Marantz, 1993 ; Fradin, 2003 ; Bauer, 1997). Les principes qui la sous-tendent lui confèrent une organisation originale : une entrée de Démonette2 correspond en effet à une relation morphologique dérivationnelle entre deux lexèmes. La BDM est ainsi conforme aux hypothèses théoriques qui considèrent que le lexème est l'unité morphologique fondamentale et que la construction dérivationnelle remplit deux fonctions : (1) créer de nouveaux lexèmes et (2) établir des relations de motivation sémantique et formelle entre les lexèmes présents dans le lexique. Par exemple, l'entrée connectant NATION$_N$ à INTERNATIONAL$_A$ rend compte du fait que l'on peut motiver le second relativement au premier : des *relations internationales* sont des *relations entre plusieurs nations*. Chaque entrée de Démonette2 comporte une description morpho-phonologique et une description morpho-sémantique (1) des lexèmes connectés ainsi que (2) de la relation qui les caractérise. Ces deux descriptions sont indépendantes. Les relations décrites ne sont pas limitées aux motivations classiques base → dérivé ; elles incluent aussi les relations sémantiquement motivées entre membres d'une famille dérivationnelle. Cette configuration permet le regroupement des familles en réseaux formels, sémantiques et dérivationnels, offrant à terme une démonstration à grande échelle de l'organisation paradigmatique du lexique construit.

# 3 Structure et sources d'approvisionnement de la BDM

**Architecture :** La constitution de Démonette2 est doublement hybride : (1) la base est élaborée à partir d'une compilation de ressources existantes et d'annotations nouvelles, manuelles, semi-automatiques et automatiques ; (2) elle combine des ressources de deux natures dont certaines documentent des *unités lexicales* et d'autres des *relations dérivationnelles* entre unités. De fait, Démonette2 se compose essentiellement de deux tables : une pour les lexèmes, et l'autre pour les relations entre lexèmes. Cette architecture limite en partie la redondance dans la base et améliore sa maintenabilité ; elle permet par ailleurs de planifier des campagnes d'annotation parallèles avec une souplesse relative. Ainsi, le travail coûteux d'annotation sémantique des lexèmes peut être mené de front avec un enrichissement du graphe des relations dérivationnelles ou avec une annotation des propriétés formelles de ces relations. La taille de la base contenant la *table des lexèmes* est maximale, de manière à garantir que l'ensemble des relations dérivationnelles décrites dans l'autre base fassent nécessairement références à des lexèmes qui y sont attestés. Pour assurer une couverture optimale, la table des lexèmes comporte l'ensemble des unités lexicales dont les 1 406 857 formes fléchies constituent *Glaff*[2] (obtenu à partir des données de *Wiktionnaire*). Outre la partie du discours, chaque lexème sera à terme décrit par l'ensemble des formes de son paradigme flexionnel transcrits au format SAMPA (cf. §4.1). Sa fréquence dans différentes ressources (Frantext[3], frWaC[4]…) et sa classe sémantique complètent cette description. La base

[2] Glàff : "Gros lexique à tout faire du français", http://redac.univ-tlse2.fr/lexiques/glaff/telechargement.html
[3] BD textuelles littéraires et journalistique hébergée à l'UMR 7118 l'ATILF : https://www.frantext.fr/

contenant la *table des relations* est, elle, alimentée initialement par le contenu de 6 ressources (cf. *infra*). Les informations transposées dans cette table concernent la structure morphologique des deux lexèmes impliqués dans la relation (est-il suffixé, préfixé, construit par conversion, et, le cas échéant, au moyen de quel affixe), et la nature de la relation (directe entre un dérivé et sa base, e.g. BASKETTEUR/BASKET, et indirecte quand elle connecte deux lexèmes de la même famille dérivationnelle comme DECORATEUR/DECORATION, FASCISME/FASCISTE ou AVIATEUR/AVIATION). Comme nous le montrons au §4, les descriptions pertinentes sont validées, et converties (semi-)automatiquement ou manuellement dans le format de la table. Elles sont en outre assorties de nouvelles informations, essentiellement sémantiques. La couverture de la base est enfin enrichie des relations qui complètent les familles dérivationnelles auxquelles appartiennent les relations déjà présentes. La gestion du projet repose sur un circuit de maintenance basé sur *git* qui différencie le rythme de modification des deux bases : la base des relations étant destinée à évoluer rapidement, sa modification est confiée à des éditeurs (humains ou automatiques) dont le travail est destiné à être injecté dès qu'il est terminé dans une version de développement, après contrôles automatiques de sa cohérence et validation par un responsable de tâche, suivant une grille de critères préalablement établis. À l'inverse, les changements dans la table des lexèmes suivent un rythme plus lent, et ne sont injectés dans la base qu'à l'occasion de la publication d'une nouvelle version de référence, sous le contrôle direct de l'administrateur. Il est à noter que la gestion de la base donne lieu à l'implémentation d'un ensemble d'outils (interface de soumission, interfaces de visualisation et d'édition à destination) qui seront valorisés et distribués comme une plateforme de maintenance de ressource morphologique.

**Sources** : Outre l'apport initial des 96 000 entrées de Demonette1.3, le contenu de Démonette2 sera dans un premier temps obtenu par migration de ressources dérivationnelles existantes, développées et validées par des morphologues (cf. Tab. 1). Ces ressources sont choisies pour leur disponibilité, leur complémentarité, la richesse des descriptions (annotations morphologiques et, pour la plupart d'entre elles, traits sémantiques et phonologiques). Leur traitement est échelonné en fonction du degré d'immédiateté de leur adaptation dans le format de Demonette2.

| Nom (auteur) | Convers (Tribout 2010) | Denom (Strnadová 2014) | Dimoc (Roché 2004, 2008, 2011a,b; Lignon, Roché 2011; Roché, Plénat 2012) | Mordan (Koehl, 2012) | Lexeur (Fabre *et al*, 2004) |
|---|---|---|---|---|---|
| Taille | 3 500 | 15 500 | 60 900 | 3 900 | 3 200 |
| Contenu | Convers N/V | Adjectifs dénom. | Noms construits sur N, V et A | Noms désadj. | Noms d'agent en *-eur* et base N / V |
| Ex. | CRI /CRIER | DUCAL /DUC | ALBIGEOIS /ALBI | BEAUTE/BEAU | BASKETTEUR/BASKET |

TABLE 1 : premières sources d'approvisionnement de la base Demonette2

Quand toutes les ressources auront été adaptées et transférées dans la base, celle-ci décrira 183 000 relations dérivationnelles réalisant environ 120 procédés de dérivation, par suffixation (*-ard, -ariat, -at, -âtre, -el, -aie, -iser, -erie, -esque, -esse, -eur, -eux, -iste* …), conversion et préfixation (*a-, anti-, bi-, co-, contre-, dé-, é-, extra-, hyper-, hypo-, in-, infra-, inter-* ...). Le contenu de Démonette2 ne se résume pas aux seules relations héritées de ces ressources. L'ambition du projet est de constituer (semi-)automatiquement les familles dérivationnelles des lexèmes codés au cours de l'étape de migration et décrire, comme autant de nouvelles entrées, les relations entre les membres de ces familles. Plusieurs approches sont envisagées : l'extension des régularités

[4] Échantillon de la toile, pour le domaine français, totalisant 1,6 milliard d'occurrences, cf. (Baroni *et al.*, 2009)

paradigmatiques encodées lors de la première phase, l'usage de réseaux de neurones (Cotterell *et al.*, 2017) ou l'application de l'analyse des concepts formels (Leeuwenberg *et al.*, 2015).

# 4 Premiers résultats

## 4.1 Un échantillon de Démonette2

Chaque entrée de Démonette2 décrit une relation dérivationnelle qui s'établit entre deux lexèmes (**Mot1, Mot2**) de la même famille dérivationnelle où Mot1 est considéré comme morphologiquement *motivé* par Mot2. En général, la motivation de Mot2 par Mot1 est également possible, ce qui fait que l'entrée symétrique (Mot2, Mot1) est aussi présente dans la base. Nous adoptons une conception élargie de la notion de famille dérivationnelle, où divers degrés de variation radicale sont possibles entre Mot1 et Mot2. La relation formelle peut être totalement transparente, comme avec (DANSE, DANSEUR) où le radical commun /dãs/ est immédiatement identifiable ; elle peut impliquer une allomorphie régulière qui n'empêche pas la reconnaissance du radical, comme avec la variation /ø/-/oz/ observée dans (NERVEUX, NERVOSITE) ; enfin, l'un des membres peut être construit sur le radical savant de l'autre, ce qui opacifie la relation formelle entre les deux, comme avec (SAINT-ETIENNE, STEPHANOIS). Chaque entrée est identifiée par la graphie et la catégorie grammaticale de Mot1 et Mot2 (cols 1 et 2 du Tab. 2), et inclut la structure morphologique des deux lexèmes (cols 3 à 6), les propriétés (cols 7 et 8) de leur relation dérivationnelle, et les caractéristiques des éventuelles transformations morphophonologiques qui y sont liées (cols 9 et 10). Le codage de ces alternances est (semi-)automatique : à chaque Moti est associée la transcription phonologique présente dans Glaff de chacun des membres de son paradigme flexionnel ; les transcriptions sont complétées et uniformisées au moyen de règles apprises à partir des codages phonétiques contenus dans la base "témoin" Flexique[5], dont le contenu est très fiable, mais la couverture réduite ; les variations (cols 9 et 10) résultent de la comparaison des transcriptions. Ainsi, les paires (Mot1, Mot2) sont regroupables selon leurs similarités morpho-phonologiques, et leurs familles dérivationnelles superposables en paradigmes formels : eg. (REALISER, REALISATEUR, REALISATION) et (ADMIRER, ADMIRATEUR, ADMIRATION) sont dans le même paradigme formel (X, Xatœr, Xasjõ). Une autre dimension paradigmatique, fondamentale pour expliquer l'organisation du lexique, est induite par les propriétés sémantiques des familles dérivationnelles. C'est pourquoi chaque entrée de Démonette est aussi caractérisée par un autre ensemble de traits, qui décrit son comportement sémantique.

| **Mot1** | **Mot2** | **Const.1** | **exp1** | **Const.2** | **exp2** | **Orienta-tion** | **Comple-xité** | **Série morpho-phonolo-gique** | **Alter-nance** |
|---|---|---|---|---|---|---|---|---|---|
| $BOIRE_V$ | $BUVEUSE_{Nf}$ | -- | -- | suf | euse | ascend | simple | X/Xøz | = |
| $ACTEUR_{Nm}$ | $ACTION_{Nm}$ | suf | eur | suf | ion | indirect | simple | Xtœr/Xsjõ | t/s |
| $BOIRE_V$ | $IMBUVABLE_A$ | -- | -- | pre | in | ascend | complexe | X/ɛ̃Xabl | |
| $FOIE_N$ | $HEPATIQUE_A$ | -- | -- | suf | ique | ascend | simple | fwa/epatik | NONE |

TABLE 2 : Entrées simplifiées de Demonette2 (extrait)

[5] http://www.llf.cnrs.fr/flexique-fr.php, (Bonami *et al.*, 2014)

## 4.2 Codages sémantiques

Les décisions à prendre en termes de codage des propriétés (morpho-)sémantiques sont cruciales, car elles conditionnent la structure et l'homogénéité du contenu de la base. Pour chaque entrée, l'information sémantique à définir est le produit de trois annotations qui se complètent, et que l'on ne trouve à notre connaissance dans aucune autre BDM : la classe ontologique de Mot1 et Mot2, la catégorie sémantique de la relation morphologique entre Mot1 et Mot2, et la motivation réciproque de Mot1 et Mot2, sous forme d'une paraphrase inspirées du modèle des *frame definitions* de Framenet (Fillmore *et al.*, 1998).

**Codage des unités du lexique :** Actuellement, par défaut, les verbes s'interprètent comme des *situations* et les adjectifs comme des *propriétés*. Pour le codage des noms, un jeu d'étiquettes a été adapté des *WordNet Unique Beginners,* désormais *UB* (Miller *et al.* 1990, Fellbaum 1998), ce qui permet de couvrir l'ensemble du spectre lexical et d'offrir un degré de généralité qui convient *a priori* à la description morphologique. La liste des *UB* (en gras), complétée par l'étiquette sous-spécifiée *Top* s'organise comme indiqué Fig.1. Nous avons établi des tableaux de correspondance entre les *UB* (ou leurs super-types, soulignés dans la Fig.1) et les types sémantiques proposés dans les ressources morphologiques intégrées dans Démonette2. L'*UB* Person correspond ainsi aux types "AGF" (agent féminin) et "AGM" (agent masculin) de Demonette1, aux types "Ah" (humain) et "Ahg" (gentilé) de Dimoc, etc. Sur la base de cet appariement, la plupart des noms issus des cinq ressources ont pu être munis d'une ou de plusieurs étiquettes normalisées.

Classe sous-spécifiée : *Top*

1 Situation [Situation stative [**Feeling, State, Attribute**]
Situation dynamique [**Act, Event**]]

2 Entité [**Objet** [Non Animé [**Objet Naturel**/**Artefact**]]
[Animé [**Animal**, **Person**] ]

Classes relationnelles : **Group, Part**

FIGURE 1 : Hiérarchie partielle des étiquettes ontologiques d'après *WordNet Unique Beginners*

**Codage des relations morpho-sémantiques :** La description des relations sémantiques associées aux règles morphologiques s'inspire du modèle des fonctions lexicales (Mel'čuk, 1996). Pour l'heure, seules les relations directes entre un dérivé et sa base ont été décrites, et dépendent du type ontologique attribué à Mot1 et Mot2. La caractérisation de la relation prend en compte la catégorie grammaticale de la base et de son dérivé et leur classe sémantique générale (*Situation* ou *Entité*) et se compose d'un type général (synonymie, résultatif, causatif, etc.), d'un schéma sémantique abstrait prenant en compte l'orientation de la relation entre Mot1 et Mot2, et du procédé dérivationnel impliqué. On distingue pour l'instant 20 types de relations sémantiques, cf. Tab 3.

| **Qualif. Gén. de la rel. sém.** | **Type sém. de Mot1** | **Type sém. de Mot2** | **Orien-tation** | **Type sém. de la rel.** | **Schéma sém. abstrait** | **Exemple (Mot1,Mot2)** |
|---|---|---|---|---|---|---|
| Entité-Entité | group x pers | person | descen | collectif | ensemble de N | (EQUIPIER, EQUIPE) |
| Situation-Entité | person | sit.dyn | descen | agent | ce(lui) qui V | (CHANTEUR, CHANTER) |
| | sit.stat | person | ascen | experiencer | ressentir ce que ressent N | (ADMIRER,ADMIRATEUR) |

TABLE 3 : Echantillon de relations morpho-sémantiques et de schémas sémantiques abstraits

Comme avec les séries morpho-phonologiques, les schémas sémantiques abstraits combinés aux types ontologiques participent à la structuration du lexique en permettant de constituer les entrées, et, de là, les familles dérivationnelles, en paradigmes sémantiques : (ENSEIGNER, ENSEIGNANT, ENSEIGNEMENT) et (CHANTER, CHANTEUR, CHANT) se retrouvent dans le même paradigme sémantique centré sur une *situation dynamique*, contrairement à (ADMIRER, ADMIRATEUR, ADMIRATION), où le verbe est statif.

**Paraphrases glosant les relations dérivationnelles :** Le troisième niveau d'annotation sémantique d'une entrée (Mot1, Mot2) est une paraphrase faisant intervenir Mot1 et Mot2 qui exprime la motivation réciproque de chaque lexème relativement au sens de l'autre, cf. Tab. 4. Cet énoncé définitoire est généralisé sous la forme d'une glose où $Mot_i$ est remplacé par son type sémantique.

| **(Mot1,Mot2)** | **Paraphrase concrète** | **Paraphrase abstraite** |
|---|---|---|
| enseigner, enseignant | Un enseignant$_1$ enseigne$_2$ | Un N*person* V*sit.dyn* |
| hôpital, hospitaliser | On hospitalise$_2$ qqc dans un hôpital$_1$ | On *Vsit.dyn* dans *Nartefact* |

TABLE 4 : Paraphrase de motivation sémantique de Mot1 et Mot2

A ce jour, une table des correspondances assure l'appariement des types sémantiques codés dans les ressources de la Tab. 1 avec les *Unique Beginners* de WordNet. 30% de la base Denom et 50% de Convers sont convertis au format Demonette2 suivant les principes présentés, et complétés de nouvelles relations. Enfin, la plateforme de dépôt est opérationnelle.

# 5 Conclusion

Au-delà des acteurs du TAL, destinataires naturels de la BDM, le contenu final du projet Démonext permettra aussi de répondre aux demandes émanant des chercheurs, des universitaires, des enseignants du primaire et des orthophonistes. En effet, la *recherche* récente en morphologie adopte des méthodes de modélisation quantitative. Celles-ci ont connu de grands succès, en particulier dans le domaine de la flexion, mais elles butent dans le domaine de la dérivation sur l'absence de ressources à large couverture contenant des informations morphologiques riches. Dans *l'enseignement supérieur* en morphologie, la base Demonette2 sera utilisée pour extraire des données morphologiques et les inclure dans des questions à réponses intégrées dans le cadre de MOOC. Enfin, la BDM permettra aux *enseignants* du *primaire* et aux cliniciens *orthophonistes* (qui s'intéressent depuis une 30aine d'années à la conscience morphologique, cf. Casalis *et al*. 2003) de construire et utiliser des outils d'évaluation, d'entraînement ou de remédiation ciblés sur la morphologie, et d'approfondir les connaissances sur l'acquisition de la morphologie dérivationnelle et les troubles développementaux qui y sont liés et pour le moment peu connus comparativement aux troubles en morphologie flexionnelle (Maillart 2003). L'orthophonie pourra également mettre à profit Démonette2 dans le traitement des adultes aphasiques présentant des troubles acquis du langage (Semenza et al. 1990, Pillon et al. 1991).

# Remerciements

Merci aux relecteurs anonymes pour leurs remarques qui ont permis d'améliorer sensiblement ce texte, ainsi qu'à nos ingénieurs partenaires du projet : Alexander Delaporte, Achille Falaise, Loïc Liégeois, et Alexandre Roulois, qui ont conçu et développé la plateforme de dépôt de Démonette2.

# Références

Baayen R.H., Piepenbrock R., Gulikers L. (1995). *The CELEX Lexical Database (CD-ROM)*. Philadelphia, PA : Linguistic Data Consortium, University of Pennsylvania.

Baroni M., Bernardini S., Ferraresi A., Zanchetta E. (2009). The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Language Resources and Evaluation* 43(3), 209-226.

Bauer L. (1997). Derivational Paradigms. *Yearbook of Morphology 1996*. Booij G., van Marle J. eds. Dordrecht : Kluwer, 243-256.

Bernhard D. (2009). Unsupervised Morphological Segmentation Based on Segment Predictability and Word Segments Alignment. Actes de *Morphochallenge 2006*, 19-23.

Bernhard D., Cartoni B., Tribout D. (2011). A Task-Based Evaluation of French Morphological Resources and Tools. *Linguistic Issues in Language Technology* 5(2), 1-41.

Bonami O., Caron G., Plancq C. (2014). Construction d'un lexique flexionnel phonétisé libre du français. Actes du *4ème Congrès Mondial de Linguistique Française*, 2583-2596.

Casalis S., Mathiot E., Bécavin A.-S., Colé P. (2003). Conscience mophologique chez les lecteurs tout venant et en difficultés. *Silexicales* 3, 57-66.

Cotterell R., Schütze H. (2015). Morphological word-embeddings. Actes de *The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1287-1292.

Cotterell R., Vylomova E., Khayrallah H., Kirov C., Yarowsky D. (2017). 'Paradigm Completion for Derivational Morphology'. Actes de *The 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Danemark, Association for Computational Linguistics*, 1-7.

Creutz M. (2003). 'Unsupervised Segmentation of Words Using Prior Distributions of Morph Length and Frequency'. Actes de *The 41th annual meeting of the ACL*, 280-287.

Creutz M., Lagus K. (2005). 'Inducing the Morphological Lexicon of a Natural Language from Unannotated Text'. Actes de *AKRR'05*, 106-113.

Dal G., Hathout N., Namer F. (1999). Construire un lexique dérivationnel: théorie et réalisations. Actes de *TALN-1999*, 115-124.

Fabre C., Floricic F., Hathout N. (2004). Collecte outillée pour l'analyse des emplois discordants des déverbaux en -eur. Communication présentée à *Journées d'étude sur la place des méthodes quantitatives dans le travail du linguiste*.

Fellbaum C. (ed.) (1998), *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
Goldsmith J. (2001). Unsupervised Learning of Morphology of a Natural Language. *Computational Linguistics* 27(2), 153-198.

Fellbaum C., Oherson A., Clark P.E. (2007). Putting semantics into Wordnet's "Morphosemantic" links. *Human Language Technology Challenges in the Information Society, 3d Language and Technology Conference*. Vetulani Z., Uszkoreit H. eds. Berlin : Springer Verlag, 350-358.

Fillmore C., Baker C., Lowe J. (1998). 'The Berkeley FrameNet Project'. Actes de *COLING-ACL*, 86-90.
Fradin B. (2003). *Nouvelles approches en morphologie*. Paris: Presses Universitaires de France.

Habash N., Dorr B. (2003). A Categorial Variation Database for English. Actes de *The North American Association for Computational Linguistics*, 96-102.

Halle M., Marantz A. (1993). Distributed Morphology and the Pieces of Inflection. *The View from Building 20: Essays in Linguistics in Honor of Sylvain Bromberger*. Hale K., Keyser S.J. eds. Cambridge, MA : MIT Press: 111-176.

Hathout N. (2009). Acquisition of morphological families and derivational series from a machine readable dictionary. *Selected Proceedings of the 6th Décembrettes: Morphology in Bordeaux*. Montermini F., Boyé G., Tseng J. eds. Cambridge, MA : Cascadilla Proceedings Project, 166-180.

Hathout N., Namer, F (2014a). La base lexicale Démonette : entre sémantique constructionnelle et morphologie dérivationnelle. Actes de *TALN*, 208-220.

Hathout N., Namer F. (2014b). Démonette, a French derivational morpho-semantic network. *Linguistic Issues in Language Technology* 11(5): 125-168.

Hathout N., Namer F. (2015). *La base lexicale morphologique du français Démonette1.2. Nancy - Toulouse*, https://www.ortolang.fr/#/market/lexicons/demonette et http://redac.univ-tlse2.fr/lexiques/demonette.html.

Hathout N., Namer F. (2016). Giving Lexical Resources a Second Life: Démonette, a Multi-sourced Morpho-semantic Network for French. Actes de *The Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 1084-1091.

Koehl A. (2012). *La construction morphologique des noms désadjectivaux suffixés en français*. Thèse de doctorat, Université de Lorraine.

Lafourcade M., Joubert A. (2008). JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes. Actes de *JADT'08 : Journées internationales d'Analyse statistiques des Données Textuelles*, 657-666.

Leeuwenberg A., Buzmakov A., Toussaint Y., Napoli A. (2015). Exploring Pattern Structures of Syntactic Trees for Relation Extraction. Actes de *ICFCA 2015*, 153-168.

Lignon S., Roché M. (2011).Entre histoire et morphophonologie, quelle distribution pour -éen vs -ien ?. *Des Unités Morphologiques au Lexique*. Roché M. ed. Paris : Hermès, 191-250.

Lux-Pogadalla V., Polguère A. (2011). Construction of a French Lexical Network: Methodological Issues. Actes de *First International Workshop on Lexical Resources, WoLeR 2011*, 54-61.

Maillart C. (2003). Les troubles pragmatiques chez les enfants présentant des difficultés langagières. Présentation d'une grille d'évaluation : la Children's Communication Checklist. *Cahiers de la SLBU* 13, 13-32.

Mel'čuk I. (1996). Lexical Functions: A tool for the description of lexical relations in the lexicon. *Lexical Functions in Lexicography and Natural Language Processing*, Wanner L. ed. Amsterdam: John Benjamins, 37–102.

Miller G., Beckwith R., Fellbaum C., Gross D., Miller K. (1990). WordNet: An online lexical database. *International Journal of Lexicography*, 3(4), 235-244.

Namer F. (2002). Acquisition automatique de sens à partir d'opérations morphologiques en français : études de cas. Actes de *TALN-2002*, 235-244.

Namer F. (2013). A Rule-Based Morphosemantic Analyzer for French for a Fine-Grained Semantic Annotation of Texts. Actes de *SFCM 2013*, 93-115.

Namer F., Hathout N., Lignon S. (2017). Adding morpho-phonological features into a French morpho-semantic resource: the Demonette derivational database. Actes de *The First International Workshop on Resources and Tools for Derivational Morphology (DeriMo)*, 49-61.

Pillon A., De Partz M.-P., Raison A.-M., Seron X. (1991). L'orange c'est le fruitier de l'orangine : a case of morphological impairment? *Language and Cognitives Processes* 6(2), 137-167.

Roché M. (2004). Mot construit ? Mot non construit ? Quelques réflexions à partir des derivés en -ier(e). *Verbum* 26(4), 459-480.

Roché M. (2008). Structuration du lexique et principe d'économie: le cas des ethniques. Actes du *1er Congrès Mondial de Linguistique Française*, 1559-1573.

Roché M. (2011a). Quel traitement unifié pour les dérivations en -isme et en -iste. *Des Unités Morphologiques au Lexique*. Roché M. ed. Paris : Hermès, 69-143.

Roché, M (2011b).Pression lexicale et contraintes phonologiques dans la dérivation en -aie du français *Linguistica* 51, 5-22.

Roché M., Plénat M. (2012). Tous les déverbaux en -at sont-ils des conversions du thème 13 ?. Actes du *3ème Congrès Mondial de Linguistique Française*, 1387-1405.

Semenza C., Butterworth B., Panzeri M., Ferreti T. (1990). Word formation : new evidence form aphasia. *Neuropsychologia* 28(5), 499-502.

Strnadová J. (2014). *Les réseaux adjectivaux. Sur la grammaire des adjectifs dénominaux en français*. Thèse de doctorat, Université Paris Diderot / Univerzita Karlova.

Tribout D. (2010). *Les conversions de nom à verbe et de verbe à nom en français*. Thèse de doctorat, Université Paris 7.

Zeller B., Šnajder J., Padó S. (2013). DERIVBASE: Inducing and Evaluating a Derivational Morphology Resource for German. Actes de *The 51st Annual Meeting of the Association for Computational Linguistics*, 1201-1211.

# Détecter la non-adhérence médicamenteuse dans les forums de discussion avec les méthodes de recherche d'information

Elise Bigeard[1] Natalia Grabar[1]
(1) CNRS, Univ Lille, UMR 8163 STL - Savoirs Textes Langage, F-59000 Lille, France,
bigeard@limsi.fr
natalia.grabar@univ-lille3.fr

RÉSUMÉ

Les méthodes de recherche d'information permettent d'explorer les données textuelles. Nous les exploitons pour la détection de messages avec la non-adhérence médicamenteuse dans les forums de discussion. La non-adhérence médicamenteuse correspond aux cas lorsqu'un patient ne respecte pas les indications de son médecin et modifie les prises de médicaments (augmente ou diminue les doses, par exemple). Le moteur de recherche exploité montre 0,9 de précision sur les 10 premiers résultats avec un corpus équilibré, et 0,4 avec un corpus respectant la distribution naturelle des messages, qui est très déséquilibrée en défaveur de la catégorie recherchée. La précision diminue avec l'augmentation du nombre de résultats considérés alors que le rappel augmente. Nous exploitons également le moteur de recherche sur de nouvelles données et avec des types précis de non-adhérence.

ABSTRACT

**Detect drug non-compliance in Internet fora using Information Retrieval methods**

Information retrieval methods allow to explore textual data. We exploit them for the detection of drug non-compliance in messages from discussions on Internet fora. Drug non-compliance happens in cases where a patient does not respect indications given by medical doctors during a drug intake (increases or decreases the dosage, for instance). The exploited search engine shows 0.9 precision on the first 10 results with balanced data, and 0.4 precision with unbalanced data respecting the natural distribution, which contain very few non-compliance messages. Precision decreases when more top answers are considered, while recall increases. We also exploit the search engine with new unseen data and on precise types of drug non-compliance.

MOTS-CLÉS : Recherche d'information, requêtes, forums, non-adhérence médicamenteuse.

KEYWORDS: Information Retrieval, Queries, Fora, Drug Non-compliance.

# 1 Introduction

Depuis de nombreuses années maintenant, les méthodes de recherche d'information montrent leur efficacité pour gérer, indexer et accéder au contenu de bases de données textuelles ou multimédia. De grandes variétés de contenus provenant de la langue générale ou des domaines de spécialité peuvent ainsi être prises en charge, comme par exemple : informations génomiques (Hersh *et al.*, 2007), informations orientées sur les patients (Goeuriot *et al.*, 2014), informations pour l'aide à la décicion clinique (Roberts *et al.*, 2015), microblogs (Lin *et al.*, 2014), informations chimiques (Lupu *et al.*, 2011) ou informations juridiques (Grossman *et al.*, 2011). Notre travail est positionné dans le

domaine médical et nous nous intéressons plus particulièrement à la non-adhérence médicamenteuse. On parle de non-adhérence médicamenteuse lorsqu'un patient ne respecte pas les instructions de son médecin ou de ses prescriptions. Le patient peut alors décider de changer le dosage de son médicament, refuser de le prendre, obtenir des médicaments sans la prescription appropriée ou bien faire d'autres actions potentiellement dangeureuses pour lui. Très peu de travaux se sont intéressés à cette problématique de santé publique. Parmi les travaux existants, l'accent principal est mis sur le sur-usage de certains types de médicaments, comme les médicaments psychotropes, qui est un cas spécifique de non-adhérence. Voilà quelques exemples de tels travaux : une étude non supervisée des tweets sur des usages non-médicaux de médicaments (Kalyanam *et al.*, 2017), la création d'une plateforme sémantique pour étudier le sur-usage de médicaments (Cameron *et al.*, 2013), étude de l'accoutumance aux médicaments et de leur consommation avec de l'alcool (Kornfield *et al.*, 2018). Il s'agit d'une question de recherche émergente et nous pouvons y voir au moins trois limitations actuelles : (1) les travaux existants sont essentiellement effectués sur les données en anglais ; (2) en dehors du sur-usage de médicaments, il existe de nombreuses autres situations de non-adhérence (Bigeard *et al.*, 2018) qui restent cependant plus rares et plus difficiles à recruter et à observer ; et (3) actuellement, il n'existe pas de données annotées manuellement sur cette question de recherche. Pour ces différentes raisons, les situations de non-adhérence médicamenteuse sont peu connues : les données de référence n'existent pas et les patients ne parlent pas de non-adhérence à leurs médecins.

Pour en savoir plus sur la non-adhérence et estimer sa prévalence dans la société, il est donc nécessaire d'étudier d'autres sources d'information. Comme dans la plupart des travaux existants, nous proposons d'étudier les messages écrits par les patients sur les réseaux sociaux, où ils parlent volontairement et sans effort de leur santé et de leurs pratiques (Gauducheau, 2008). Les réseaux sociaux sont en effet devenus une source d'information inestimable pour de nombreux domaines de recherche, tels que la géolocalisation, la fouille d'opinion, l'extraction d'évènement, la traduction ou encore le résumé automatique (Louis, 2016). Dans cette étude nous proposons d'identifier des situations de non-adhérence à l'aide de méthodes de recherche d'information. Notre objectif est d'identifier et d'analyser des messages relatifs à la non-adhérence provenant de forums de santé francophones.

# 2 Méthode pour détecter la non-adhérence médicamenteuse

La méthode doit prendre en charge les spécificités des données traitées : (1) cibler les cas de non-adhérence médicamenteuse, ce qui demande de pouvoir accéder aux entités pertinentes comme les noms de médicaments et les expressions de non-adhérence ; (2) prendre en charge l'écriture non normée des réseaux sociaux ; et (3) prendre en compte le faible volume de données annotées disponibles et la faible prévalence du phénomène de non-adhérence dans les réseaux sociaux. Ce dernier point correspond à la principale motivation d'exploiter les méthodes de recherche d'information car la détection de messages avec la non-adhérence peut alors être effectuée de manière non-supervisée.

## 2.1 Corpus et données de référence

Notre corpus est constitué de messages collectés sur plusieurs forums de santé francophones :

— Doctissimo [1] est un forum de santé très populaire. Cette plateforme permet aux patients et à leurs proches de discuter des maladies, des médicaments et de la vie quotidienne. Nous avons

1. `http://forum.doctissimo.fr`

collecté des messages dans plusieurs sous-forums (grossesse, médicaments, douleurs de dos, accidents sportifs, diabète). Les messages collectés ont été postés entre 2010 et 2017 ;
— AlloDocteur [2] est une plateforme de questions/réponses où les questions des patients sont traitées par des médecins ;
— masante.net [3] est une plateforme de questions/réponses avec des réponses de médecins ;
— Les diabétiques [4] est un forum spécialisé dans le diabète.

Dans tous ces forums, les contributeurs sont principalement des patients et leurs proches qui s'adressent à la communauté pour poser des questions ou pour témoigner de leur expérience avec leurs maladies et le processus de soins (les médicaments pris, l'évolution de la maladie, etc.).

Ce corpus est pré-traité : (1) les messages de plus de 2 500 caractères sont exclus car leur contenu hétérogène est difficile à analyser et catégoriser, pour les annotateurs comme pour les classifieurs ; (2) nous annotons les messages automatiquement avec les noms de médicaments grâce à un vocabulaire construit à partir de plusieurs sources existantes et leur association avec les codes ATC correspondants (Skrbo *et al.*, 2004). (3) Cela permet de sélectionner les messages contenant au moins un nom de médicament, pour un total de 119 562 messages (soit plus de 15,5M de mots).

Pour créer les données de référence, un sous-ensemble contenant 1 850 messages est annoté manuellement. Les annotateurs assignent à chaque message l'une des deux catégories suivantes :
— *non-adhérence* : les messages parlent de non-adhérence. Pour cette catégorie, l'annotateur doit également fournir une courte description de ce sur quoi porte la non-adhérence (sur-usage, changement de dose...) en texte libre. Par exemple, ce message correspond à un oubli de prise : *"bon moi la miss boulette et la tete en l'air je devais commencer mon "utrogestran 200" a j16 bien sur j'ai oublier ! donc je l'ai pris ce soir ! ! ! !"*
— *adhérence* : les messages contiennent un usage normal de médicament (*"Mais la question que je pose est 'est ce que c'est normal que le loxapac que je prends met des heures à agir ? ? ?"* ou pas d'usage (*"ouf boo, repose toi surtout, il ne t'a pas prescris d'aspegic nourisson ? ?"*)

Les annotateurs ont aussi la possibilité d'indiquer que la catégorie d'un message est indécidable. Dans ce cas, et en cas de désaccord entre annotateurs, l'annotation finale est décidée ultérieurement suite à un consensus. L'annotation est effectuée par trois annotateurs : un expert en pharmacologie et deux informaticiens familiers avec le domaine médical et les tâches d'annotation. Au total, 1 850 messages sont doublement annotés : 1 717 messages sont attribués à la catégorie *adhérence* et 133 messages à la catégorie *non-adhérence*. Cela indique clairement que les catégories sont naturellement déséquilibrées. L'accord inter-annotateur (Cohen, 1960) est de 0,46 : il s'agit d'un accord modéré (Landis & Koch, 1977). Dans certains messages, l'informations est incomplète ou ambiguë, ce qui complique l'annotation : le patient peut déclarer l'arrêt du médicament sans préciser s'il s'agit de sa propre décision ou s'il a obtenu l'accord de son médecin. Notons aussi que la catégorie *non-adhérence* contient 16 types, contenant 1 à 29 messages. Par exemple *perte/prise de poids* contient 2 messages, *usage récréatif* contient 2 messages, *tentative de suicide* 2 messages, et *sur-usage* 20 messages.

## 2.2 Exploitation de méthodes de recherche d'information

Le corpus est segmenté, annoté en parties du discours et lemmatisé avec Treetagger (Schmid, 1994). Suite à ces traitements, trois versions des messages sont disponibles : *formes* où les messages sont

2. http://www.allodocteur.fr
3. http://ma-sante.net
4. http://www.lesdiabetiques.com

seulement segmentés et la casse mise en minuscules ; *lemmes* où les messages sont de plus lemmatisés, les nombres remplacés par une marque de substitution unique, et les diacritiques supprimées ; *lemmes lexicaux* où seuls les verbes, noms, adjectifs et adverbes sont conservés. Nous exploitons ensuite le système de recherche d'information Indri (Strohman *et al.*, 2005) pour détecter les messages de non-adhérence. Les données de référence permettent d'effectuer deux expériences, selon que la distribution des messages avec les cas de non-adhérence est équilibrée ou naturelle :

— *distribution équilibrée* : la proportion de messages d'adhérence et de non-adhérence est équilibrée dans l'ensemble d'évaluation, qui contient alors 44 messages d'adhérence et 44 messages de non-adhérence, pour un total de 88 messages. Les 44 messages de non-adhérence restants sont utilisés pour construire la requête ;

— *distribution naturelle* : la proportion de non-adhérence respecte la distribution naturellement observée dans le corpus. L'ensemble d'évaluation est alors constitué de 42 messages de non-adhérence et de 558 messages d'adhérence, pour un total de 600 messages. Les 42 messages de non-adhérence restants sont utilisés pour construire la requête.

Deux méthodes permettent de faire les requêtes et prendre en compte les spécificités de l'écriture :

R1 : Les messages sont segmentés en mots et convertis en lexique où chaque mot est associé au poids correspondant à sa fréquence dans ces messages. Lors de la recherche, Indri ordonne les résultats selon leur ressemblance avec les termes de non-adhérence donnés dans la requête ;

R2 : Nous entraînons des plongements de mots (Mikolov *et al.*, 2013a,b) sur un corpus de 20 000 messages non-annotés. Ensuite, pour chaque mot lexical, nous retenons ses 10 mots les plus proches. Lors de la construction de la requête, les poids des mots prennent en compte les plongements lexicaux : chaque fois qu'une occurence d'un mot est comptabilisée, ses 10 mots les plus proches se voient attribuer une occurence supplémentaire également.

Pour l'évaluation, nous calculons la précision, le rappel et la F-mesure parmi les $N$ premiers résultats retournés par Indri. Enfin, nous comparons les résultats à ceux obtenus par apprentissage supervisé avec NaiveBayes (John & Langley, 1995). Les ensembles d'évaluation sont identiques. En ce qui concerne les ensembles d'entraînement, nous ajoutons 44 messages d'adhérence pour la distribution équilibrée et 558 pour la distribution naturelle afin de construire les requêtes Indri. Ainsi la proportion de messages de chaque classe est identique dans les ensembles d'entrainement et d'évaluation. Comme indiqué, au sein de la catégorie de non-adhérence, il peut en exister des types plus précis. L'objectif consiste alors à détecter des messages appartenant à un type spécifique de non-adhérence mais en mode prospectif. Nous exploitons alors les noms de la non-adhérence (*sur-dosage, alcool, tentative de suicide...*) et notre connaissance du corpus : les requêtes sont appliquées à un corpus de 20 000 messages sélectionnés aléatoirement. Ces résultats sont évalués en calculant la précision.

# 3 Résultats et discussion

Les résultats obtenus avec la requête $R1$ sont présentés dans le tableau 1. L'évaluation est effectuée pour les 10, 50 et 100 premiers résultats retournés par Indri et en fonction de la distribution équilibrée ou naturelle des données. Les meilleurs résultats sont toujours obtenus avec la version *lemmes lexicaux*, ce qui démontre que la lemmatisation améliore les résultats (précision et rappel) et que les mots lexicaux apportent moins de bruit par rapport aux mots grammaticaux et aux formes. Le rappel augmente avec l'augmentation du nombre de résultats pris en compte car cela permet de se rapprocher de l'ensemble des messages contenant les cas de non-adhérence (42 et 44), alors que la précision diminue. Avec la distribution équilibrée, la précision la plus faible, obtenue avec les 10 premiers

| | *Distribution équilibrée* | | | *Distribution naturelle* | | |
|---|---|---|---|---|---|---|
| | *Précision* | *Rappel* | *F-mesure* | *Précision* | *Rappel* | *F-mesure* |
| | *10 premiers résultats* | | | | | |
| *formes* | 0,500 | 0,114 | 0,186 | 0,100 | 0,024 | 0,038 |
| *lemmes* | 0,800 | 0,182 | 0,296 | 0,200 | 0,047 | 0,076 |
| *lemmes lexicaux* | **0,900** | 0,204 | 0,333 | **0,400** | 0,095 | 0,153 |
| | *50 premiers résultats* | | | | | |
| *formes* | 0,560 | **0,636** | **0,596** | 0,120 | 0,143 | 0,130 |
| *lemmes* | 0,540 | 0,610 | 0,574 | 0,140 | 0,167 | 0,152 |
| *lemmes lexicaux* | 0,560 | **0,636** | **0,596** | 0,220 | 0,262 | **0,239** |
| | *100 premiers résultats* | | | | | |
| *formes* | 0,500 | 1,000 | 0,666 | 0,120 | 0,286 | 0,169 |
| *lemmes* | 0,500 | 1,000 | 0,666 | 0,120 | 0,286 | 0,169 |
| *lemmes lexicaux* | 0,500 | 1,000 | 0,666 | 0,150 | **0,357** | 0,211 |

TABLE 1 – Résultats du système de recherche d'information pour la catégorisation des messages dans la classe *non-adhérence*. Le meilleur résultat pour chaque métrique est mis en gras.

résultats sur le corpus *formes*, est de 0,5. Elle est nettement améliorée avec les autres versions du corpus (*lemmes* et *lemmes lexicaux*), de même que lorsque les 50 premières réponses sont prises en compte. Elle est maintenue avec les 100 premiers résultats. Cela indique que le système parvient à bien différencier les cas de non-adhérence. Les meilleurs résultats sont obtenus avec les 100 premiers messages, ce qui privilégie le rappel tout en maintenant la précision. Avec la distribution naturelle de la non-adhérence, la tâche devient plus complexe et les performances restent faibles.

| | *Distribution équilibrée* | | | *Distribution naturelle* | | |
|---|---|---|---|---|---|---|
| | *Précision* | *Rappel* | *F-mesure* | *Précision* | *Rappel* | *F-mesure* |
| *formes* | 0,630 | 0,773 | 0,694 | 0,167 | **0,524** | 0,253 |
| *lemmes* | 0,618 | 0,773 | 0,687 | 0,169 | **0,524** | 0,256 |
| *lemmes lexicaux* | **0,636** | **0,795** | **0,707** | **0,186** | 0,500 | **0,271** |

TABLE 2 – Résultats de NaiveBayes pour la catégorisation des messages dans la classe *non-adhérence*. Le meilleur résultat pour chaque métrique est mis en gras.

Le tableau 2 présente les résultats obtenus avec le classifieur supervisé NaiveBayes entraîné et testé sur les mêmes données. Sur les 10 premiers résultats, Indri obtient une meilleure précision que Bayes. En revanche Bayes atteint un meilleur rappel et F-mesure. Cela indique que les méthodes de recherche d'information peuvent être utiles pour sélectionner les messages les plus susceptibles de correspondre à la non-adhérence médicamenteuse. Une analyse des erreurs indique que souvent ce sont les expressions d'excès qui apportent la confusion dans les résultats : le moteur de recherche, de même que le classifieur, peuvent classer d'autres types de messages dans la catégorie de non-adhérence.

La requête $R2$, qui prend en compte les vecteurs de mots issus des plongements lexicaux, détériore systématiquement les résultats. À titre d'exemple, avec les 100 premiers résultats et les données équilibrées, la précision diminue à 0,44, alors qu'avec les données non équilibrées et la version du corpus *lemmes lexicaux*, la F-mesure devient 0,013 (0,007 de précision et 0,167 de rappel). Notons que les résultats avec les versions *formes* des corpus sont très faiblement impactés. D'autres expériences sont cependant nécessaires pour évaluer plus précisément le rôle et le poids des mots-clés sur les

résultats. La limitation principale de cette série d'expériences est la petite quantité de données de référence disponible. La généralisation des résultats est donc limitée.

Dans les expériences qui suivent, nous utilisons le moteur de recherche Indri pour obtenir davantage d'exemples pour les types de non-adhérence plus précis et disposant de peu d'exemples actuellement. Le moteur de recherche est exploité avec ses paramètres par défaut. Nous retenons et analysons les 20 premiers résultats. Plusieurs types de non-adhérences sont ainsi traités. Comme indiqué, les noms de types de non-adhérences sont essentiellement utilisés pour construire ces requêtes.

**Gain et perte de poids.** Les mots-clés de la requête (*poids, kilo, grossir, maigrir*) sont appliqués au corpus lemmatisé. Nous recherchons avec cette requête des messages parlant de médicaments pris dans l'objectif de perdre du poids, ainsi que des messages parlant de perte ou de gain de poids comme effet secondaire d'un médicament. Parmi les 20 premiers messages, un seul message de non-adhérence est trouvé. D'autres réponses parlent de choses proches (essentiellement, le changement de poids comme effet indésirable de médicament). Cela peut être un reflet de la réalité, auquel cas le mésusage de médicament pour perdre du poids est bien moins fréquent que les changements de poids comme effet secondaire. Ce peut aussi être un artefact du corpus utilisé, où de nombreux messages parlent d'anti-dépresseurs, qui sont une classe de médicaments avec lesquels les changements de poids sont un effet secondaire fréquent. Cette requête nous donne 0.05 de précision.

**Usage récréatif.** Ici nous cherchons à trouver des messages où les médicaments sont utilisés pour un usage récréatif : pour se "droguer", "planer", provoquer des hallucinations, etc. Nous avons testé plusieurs requêtes :

- Les mots-clés *drogue, droguer* semblent être de bons candidats car, dans notre corpus, les patients les utilisent en référence à des médicaments de type neuroleptique, afin d'illustrer leur sentiment que l'effet de ces médicaments, ainsi que leur risque de dépendance, sont similaires à ceux des drogues. Ces effets sont vécus négativement et ne sont pas recherchés. Nous trouvons par exemple les messages de type *J'ai été drogué pendant 3 ans au xanax* ou *Sa soulage mais ses une vrai drogue ce truc ! ! !*. Cette requête trouve 15 résultats pertinents parmi les 20 analysés ;
- Les mots-clés *hallu, allu, hallucination* fournissent 2 messages de non-adhérence et plusieurs messages avec des contenus proches (7 messages avec l'hallucination comme effet indésirable non-recherché, 11 messages où les patients souffrent d'hallucinations) ;
- Le mot-clé *planer* fournit 9 messages attendu de type *J'ai déjà posté quelques sujets à propos de ce fléau qu'est le stilnox (...) je prends du stilnox, pour m'évader, pour planer"*. D'autres messages trouvés sont liés à l'effet de "planer" provoqué par un médicament mais sans être recherché par le patient.

Ces différentes requêtes donnent en moyenne 0,43 de précision.

**Suicide.** La requête contient uniquement le mot-clé *suicide*. Le but de cette requête est de trouver des messages parlant de tentatives de suicide par ingestion de médicaments. Cette requête montre une précision de 0,3, avec 5 messages sur les médicaments pouvant augmenter le risque de suicide et un message où l'auteur raconte une tentative de suicide provoquée par un sevrage médicamenteux qui s'est très mal passé. D'autres réponses ne sont pas correctes : les médicaments et le suicide ne sont pas liés entre eux, une critique du suicide, etc.

**Sur-usage.** On parle de sur-usage lorsque le patient consomme une plus grande dose de médicaments par rapport à ce qu'indique son ordonnance ou par rapport à la dose maximale authorisée. Le mot-clé *boites* (au pluriel) et le corpus *formes* sont utilisés pour cette requête. Dans notre corpus, ce mot-clé est en effet souvent lié à la quantification de médicaments. Cette requête donne 0.65 de précision.

D'autres messages retrouvés parlent de tentatives de suicide médicamenteuses, de propositions de donner les boites de médicaments non utilisées, ou de sujets sans lien avec la requête.

**Alcool.** Le mot-clé *alcool* montre 0,6 de précision : 12 messages parlent en effet d'interactions entre alcool et médicaments. D'autres messages concernent les médicaments prescrits pour le sevrage alcoolique ou bien n'ont pas de lien avec le sujet. De plus, nous remarquons que parmi les 12 messages parlant d'interactions entre médicament et alcool, 8 messages concernent les neuroleptiques. Cette classe de médicaments est en effet connue pour ses interactions avec l'alcool, ce qui peut expliquer leur présence dans les résultats.

Globalement, nous pouvons voir que les requêtes ponctuelles concernant les types précis de non-adhérence, qui ont de très faibles effectifs dans les données de référence constituées auparavant (souvent pas plus de 3 messages), apportent des résultats intéressants avec une précision moyenne de 0,40. Ce type d'interrogation de corpus permet en effet d'augmenter les données de référence. Des travaux futurs sont nécessaires pour automatiser la création de requêtes pour la détection et le classement de types de messages avec de très faibles effectifs.

# 4 Conclusion

Dans cette étude, nous avons présenté l'utilisation de méthodes de recherche d'information pour détecter les cas de non-adhérence médicamenteuse dans les forums de discussion. Dans les données de référence, les messages sont annotés manuellement en deux catégories : *adhérence* et *non-adhérence*. Nous utilisons le système de recherche d'information Indri pour détecter les messages de non-adhérence. Deux ensembles de requêtes sont construits, avec et sans les informations provenant des plongements lexicaux. Sans les plongements lexicaux, Indri obtient 0,9 de précision sur les 10 premiers résultats avec un corpus équilibré, et 0,4 avec un corpus respectant la distribution naturelle des messages avec la non-adhérence. La distribution est alors très déséquilibrée avec très peu de cas de non-adhérence. La précision diminue avec l'augmentation du nombre de résultats considérés alors que le rappel augmente. La meilleure F-mesure (0,666) est obtenue avec 100 réponses et les données équilibrées. Le corpus *lemmes lexicaux* fournit de meilleurs résultats. Avec les plongements lexicaux, les résultats sont déteriorés. Seules les versions *formes* des corpus sont pas ou très faiblement impactées. Le moteur de recherche est ensuite utilisé pour détecter des types de non-adhérence plus précis : gain et perte de poids, usage récréatif de médicaments, tentative de suicide, sur-usage de médicaments et consommation d'alcool avec les médicaments. Ces types de non-adhérence ont un très faible nombre de messages dans les données de référence. À cette étape, nous obtenons une précision moyenne de 0,40, ce qui indique que nous pouvons détecter de nouveaux messages pertinents. Dans de futurs travaux, les méthodes de requêtage pourront être améliorées et mieux automatisées.

# Remerciements

Ce travail fait partie du projet *MIAM (Maladies, Interactions Alimentation-Médicaments)* financé par l'ANR sous la référence ANR-16-CE23-0012. Ce travail s'inscrit également dans le programme *Drugs Systematized Assessment in real-liFe Environnement (DRUGS-SAFE)* financé par l'Agence Nationale de Sécurité du Médicament et des Produits de Santé. Cette publication ne représente pas nécessairement l'opinion de l'ANSM.

# Références

BIGEARD E., GRABAR N. & THIESSARD F. (2018). Typology of drug misuse created from information available in health fora. In *MIE 2018*, p. 1–5.

CAMERON D., SMITH G. A., DANIULAITYTE R., SHETH A. P., DAVE D., CHEN L., ANAND G., CARLSON R., WATKINS K. Z. & FALCK R. (2013). PREDOSE : a semantic web platform for drug abuse epidemiology using social media. **46**(6), 985–997.

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1), 37–46.

GAUDUCHEAU N. (2008). La communication des émotions dans les échanges médiatisés par ordinateur : bilan et perspectives. *Bulletin de psychologie*, p. 389–404.

GOEURIOT L., KELLY L., LI W., PALOTTI J., PECINA P., ZUCCON G., HANBURY A., JONES G. & MÜLLER H. (2014). Share/clef ehealth evaluation lab 2014, task 3 : User-centred health information retrieval. In *CLEF*, Lecture Notes in Computer Science (LNCS), p. 43–61 : Springer.

GROSSMAN M. R., CORMACK G. V., HEDIN B. & OARD D. W. (2011). Overview of the trec 2011 legal track. In E. M. VOORHEES & A. ELLIS, Eds., *The Twentieth Text REtrieval Conference Proceedings (TREC 2011)*.

HERSH W., COHEN A. & ROBERTS P. (2007). *TREC 2007 Genomics Track Overview*. Rapport interne, TREC Genomics.

JOHN G. H. & LANGLEY P. (1995). Estimating continuous distributions in bayesian classifiers. In M. KAUFMANN, Ed., *Eleventh Conference on Uncertainty in Artificial Intelligence*, p. 338–345, San Mateo.

KALYANAM J., KATSUKI T., LANCKRIET G. R. G. & MACKEY T. K. (2017). Exploring trends of nonmedical use of prescription drugs and polydrug abuse in the twittersphere using unsupervised machine learning. *Addictive Behaviors*, **65**, 289–295.

KORNFIELD R., SARMA P. K., SHAH D. V., MCTAVISH F., LANDUCCI G., PE-ROMASHKO K. & GUSTAFSON D. H. (2018). Detecting recovery problems just in time : Application of automated linguistic analysis and supervised machine learning to an online substance abuse forum. *J Med Internet Res*, **20**(6), 1–17.

LANDIS J. & KOCH G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.

LIN J., WANG Y., EFRON M. & SHERMAN G. (2014). Overview of the trec-2014 microblog track. In E. M. VOORHEES & A. ELLIS, Eds., *The Twenty-Third Text REtrieval Conference Proceedings (TREC 2014)*.

LOUIS A. (2016). Natural language processing for social media. *Computational Linguistics*, **42**(4), 833–836.

LUPU M., JIASHU Z., HUANG J., GURULINGAPPA H., FLUCK J., ZIMMERMAN M., FILIPPOV I. & TAIT J. (2011). Overview of the trec 2011 chemical ir track. In E. M. VOORHEES & A. ELLIS, Eds., *The Twentieth Text REtrieval Conference Proceedings (TREC 2011)*.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013a). Efficient estimation of word representations in vector space. In *Workshop at ICLR*.

MIKOLOV T., SUSTKEVER I., CHEN K., CORRADO G. & DEAN J. (2013b). Distributed representations of words and phrases and their compositionality. In *NIPS*.

ROBERTS K., SIMPSON M. S., VOORHEES E. M. & HERSH W. R. (2015). Overview of the trec 2015 clinical decision support track. In E. M. VOORHEES & A. ELLIS, Eds., *The Twenty-Fourth Text REtrieval Conference Proceedings (TREC 2015)*.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *ICNMLP*, p. 44–49, Manchester, UK.

SKRBO A., BEGOVIĆ B. & SKRBO S. (2004). Classification of drugs using the atc system (anatomic, therapeutic, chemical classification) and the latest changes. *Med Arh*, **58**(2), 138–41.

STROHMAN T., METZLER D., TURTLE H. & CROFT W. B. (2005). Indri : a language-model based search engine for complex queries. In *Proceedings of the International Conference on Intelligent Analysis*.

# Détection automatique de phrases parallèles dans un corpus biomédical comparable technique/simplifié

Rémi Cardon Natalia Grabar
CNRS, UMR 8163, F-59000 Lille, France
Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France
remi.cardon@univ-lille.fr, Natalia.grabar@univ-lille.fr

## RÉSUMÉ

Les phrases parallèles contiennent des informations identiques ou très proches sémantiquement et offrent des indications importantes sur le fonctionnement de la langue. Lorsque les phrases sont différenciées par leur registre (comme expert *vs.* non-expert), elles peuvent être exploitées pour la simplification automatique de textes. Le but de la simplification automatique est d'améliorer la compréhension de textes. Par exemple, dans le domaine biomédical, la simplification peut permettre aux patients de mieux comprendre les textes relatifs à leur santé. Il existe cependant très peu de ressources pour la simplification en français. Nous proposons donc d'exploiter des corpus comparables, différenciés par leur technicité, pour y détecter des phrases parallèles et les aligner. Les données de référence sont créées manuellement et montrent un accord inter-annotateur de 0,76. Nous expérimentons sur des données équilibrées et déséquilibrées. La F-mesure sur les données équilibrées atteint jusqu'à 0,94. Sur les données déséquilibrées, les résultats sont plus faibles (jusqu'à 0,92 de F-mesure) mais restent compétitifs lorsque les modèles sont entraînés sur les données équilibrées.

## ABSTRACT

**Automatic detection of parallel sentences in comparable biomedical corpora**

Parallel sentences provide identical or semantically similar information which gives important clues on language. When sentences vary by their register (like expert *vs* non-expert), they can be exploited for the automatic text simplification. The aim of text simplification is to improve the understanding of texts. For instance, in the biomedical field, simplification may permit patients to understand better medical texts in relation to their health. Yet, there is currently very few resources for the simplification of French texts. We propose to exploit comparable corpora, which are distinguished by their technicality, to detect parallel sentences and to align them. The reference data are created manually and show 0.76 inter-annotator agreement. We perform experiments on balanced and imbalanced data. The results on balanced data reach up to 0.94 F-measure. On imbalanced data, the results are lower (up to 0.92 F-measure) but remain competitive when using classification models trained on balanced data.

MOTS-CLÉS : Simplification, classification, similarité, phrases parallèles, corpus comparables, domaine médical.

KEYWORDS: Simplification, classification, similarity, parallel sentences, comparable corpora, medical domain.

# 1 Introduction

Les phrases parallèles possèdent une sémantique similaire et la véhiculent d'une manière qui peut varier selon un axe donné. Typiquement, les phrases parallèles sont collectées dans deux langues différentes et correspondent à des traductions. Dans la langue générale, le corpus Europarl (Koehn, 2005) contient de telles phrases dans plusieurs paires de langues. Cependant, l'axe à partir duquel on observe le parallélisme peut se trouver à d'autres niveaux, comme le registre de langue expert *vs.* non-expert. La paire de phrases ci-après illustre la différence de technicité entre deux phrases :

— Expert : *Les médicaments inhibant le péristaltisme sont contre-indiqués dans cette situation*
— Non-expert : *Dans ce cas, ne prenez pas de médicaments destinés à bloquer ou ralentir le transit intestinal*

Les paires de phrases parallèles fournissent des informations utiles sur le lexique, les structures syntaxiques, les traits stylistiques, etc. propres à un registre donné. Ainsi, les paires collectées dans différentes langues servent à la traduction automatique, alors que les paires ayant une technicité différente peuvent servir à la simplification automatique. L'objectif de la simplification consiste à produire une version simplifiée d'un texte, afin d'éliminer, de restructurer ou de remplacer les segments difficiles, tout en maintenant le sens. La simplification peut s'occuper de différents aspects, comme le lexique, la syntaxe, la sémantique, la pragmatique ou encore la structure d'un document.

La simplification automatique de textes peut être une étape de pré-traitement pour les applications de TAL ou pour produire des versions adaptées de textes pour différents utilisateurs humains. Par exemple, les documents simplifiés peuvent être destinés aux enfants (Vu *et al.*, 2014), aux étrangers ou personnes mal alphabétisées (Paetzold & Specia, 2016), aux personnes souffrant de pathologies mentales ou neurodégénératives (Chen *et al.*, 2016), ou au grand public qui cherche à comprendre des documents spécialisés (Leroy *et al.*, 2013). Notre objectif est de préparer les ressources nécessaires pour la création de documents médicaux simplifiés pour le grand public. Il a en effet été montré que les documents médicaux sont souvent difficiles à comprendre par les patients et leurs familles (AMA, 1999; Mcgray, 2005; Rudd, 2013), ce qui peut avoir une influence négative sur le processus de soins. Plus particulièrement, nous proposons de détecter des phrases parallèles et de les aligner. Comme il n'existe pas de textes parallèles en français qui soient différenciés par leur technicité, nous proposons d'exploiter le corpus comparable CLEAR (Grabar & Cardon, 2018), où les textes traitent des mêmes sujets mais diffèrent par leur technicité. Nous étudions également l'influence du déséquilibre sur la difficulté de la tâche, qui est en effet une caractéristique naturelle des données textuelles.

Il existe plusieurs travaux en détection et alignement de phrases parallèles au sein de corpus comparables bilingues pour les besoins de la traduction automatique. Différentes méthodes sont exploitées pour cela, comme des systèmes de recherche d'information cross-langue (Utiyama & Isahara, 2003; Munteanu & Marcu, 2006), des arbres d'alignement de séquences (Munteanu & Marcu, 2002) ou des traductions automatiques mutuelles (Yang & Li, 2003; Munteanu & Marcu, 2005; Kumano *et al.*, 2007; Abdul-Rauf & Schwenk, 2009). Souvent, il est nécessaire d'effectuer ensuite un filtrage pour la sélection de bonnes propositions. En ce qui concerne les travaux en alignement de phrases parallèles dans les corpus comparables monolingues, la difficulté principale est liée au faible chevauchement lexical entre les phrases. Récemment, cette tâche a gagné en popularité autour de la langue générale grâce aux tâches STS (semantic text similarity) et les compétitions *SemEval* (Agirre *et al.*, 2013, 2015, 2016) : pour une paire de phrases donnée, l'objectif consiste à prédire leur niveau de similarité sémantique et d'y attribuer un score allant de 0 (sémantique indépendante) à 5 (identité sémantique). Plusieurs types de méthodes sont proposés :

— Les *méthodes lexicales* se basent sur la similarité des mots ou des segments sublexicaux

(Madnani *et al.*, 2012). Nous trouvons parmi les descripteurs exploités : chevauchement lexical, longueur des phrases, distance d'édition des chaînes de caractères, nombres, entités nommées, la sous-chaîne commune la plus longue (Clough *et al.*, 2002; Zhang & Patrick, 2005; Qiu *et al.*, 2006; Nelken & Shieber, 2006; Zhu *et al.*, 2010) ;

— Les *méthodes basées sur des connaissances externes* utilisent des ressources comme WordNet (Miller *et al.*, 1993) ou PPDB (Ganitkevitch *et al.*, 2013). Parmi les descripteurs exploités se trouvent : recoupement avec des ressources externes, distance et intersection entre synsets, similarité sémantique entre les graphes, présence de synonymes, hyperonymes ou antonymes (Mihalcea *et al.*, 2006; Fernando & Stevenson, 2008; Lai & Hockenmaier, 2014) ;

— Les *méthodes basées sur la syntaxe* exploitent la modélisation syntaxique des phrases. Les descripteurs utilisés sont : catégories morphosyntaxiques, chevauchement syntaxique, dépendances syntaxiques, constituants, relations prédicatives, distance d'édition entre arbres syntaxiques (Wan *et al.*, 2006; Severyn *et al.*, 2013; Tai *et al.*, 2015; Tsubaki *et al.*, 2016) ;

— Les *méthodes basées sur les corpus* exploitent par exemple des méthodes distributionnelles, la LSA et les plongements lexicaux (Barzilay & Elhadad, 2003; Guo & Diab, 2012; Zhao *et al.*, 2014; Kiros *et al.*, 2015; He *et al.*, 2015; Mueller & Thyagarajan, 2016).

À notre connaissance, il n'existe pas de travaux en détection et alignement de phrases parallèles en domaine spécialisé, comme le domaine biomédical. Nous pensons que cela peut rendre la tâche d'alignement de phrases plus compliquée parce qu'il existe une variation lexicale importante entre le registre technique et simplifié, comme on peut le voir dans les exemples présentés au début de cette section. Dans ce qui suit, nous présentons d'abord les données linguistiques utilisées et les méthodes proposées. Ensuite nous présentons et discutons les résultats obtenus, avant de conclure avec des perspectives sur le travail à venir.

## 2 Données linguistiques

Nous utilisons le corpus comparable médical CLEAR disponible en ligne [1] qui contient trois sous-corpus comparables en français. Les documents de ces sous-corpus sont regroupés par paires : les textes de chaque paire traitent du même sujet mais varient par leur degré de technicité. Trois genres sont représentés : l'information sur les médicaments, les résumés de la littérature scientifique médicale et des articles encyclopédiques. Au total, ce corpus contient 16 190 paires de documents, avec plus de 15M d'occurrences de mots dans la version technique et 35M d'occurrences dans la version simplifiée.

Les données de référence sont créées manuellement à partir de 39 paires de textes sélectionnés aléatoirement. La référence contient les paires de phrases alignées, qui associent les contenus techniques et simplifiés. L'alignement est effectué par deux annotateurs indépendants selon ces critères :

1. exclure les paires de phrases identiques ou variant par la ponctuation ou mots grammaticaux ;
2. inclure les paires de phrases avec des variations morphologiques, comme dans : *Ne pas dépasser la posologie recommandée.* et *Ne dépassez pas la posologie recommandée.* ;
3. inclure les paires de phrases avec une sémantique équivalente, comme dans : *Les médicaments inhibant le péristaltisme sont contre-indiqués dans cette situation.* et *Dans ce cas, ne prenez pas de médicaments destinés à bloquer ou ralentir le transit intestinal.* ;
4. inclure les paires de phrases où une phrase est comprise dans l'autre, ce qui permet d'associer plusieurs phrases à une seule, comme dans : *C'est un organe fait de tissus membraneux et*

1. http://natalia.grabar.free.fr/resources.php#clear

*musculaires, d'environ 10 à 15 mm de long, qui pend à la partie moyenne du voile du palais.* et *Elle est constituée d' un tissu membraneux et musculaire.* ;

5. exclure les paires de phrases avec intersection sémantique, où chaque phrase contient de l'information qui lui est propre, en plus de la sémantique commune. Si l'accès à l'information commune aux deux phrases est intéressant en soi, cela complexifie grandement la tâche de reconnaissance des phrases parallèles. Par ailleurs, il devient plus difficile d'exploiter ce type de paires de phrases pour la préparation du lexique et de règles de simplification.

Suite au consensus, 663 paires de phrases alignées ont été produites. L'accord inter-annotateur est de 0,76 (Cohen, 1960). Le tableau 1 présente les données de référence obtenues : le nombre de documents, de phrases et d'occurrences de mots pour les sous-corpus et les registres technique et simplifié, de même que le taux d'alignement entre ces deux registres. Ces informations sont détaillées pour chaque sous-corpus : informations sur les médicaments (*Méd.*), littérature scientifique (*Sci.*) et articles encyclopédiques (*Wiki.*). Le taux d'alignement est indiqué car, dû aux principes de création des versions simplifiées, ces trois corpus ne montrent pas la même capacité à produire des paires de phrases parallèles. Le taux d'alignement correspond au rapport entre le nombre de phrases alignées et le nombre total de phrases dans un corpus donné. De manière peu surprenante, seule une petite fraction de toutes les paires possibles peut être alignée.

TABLE 1 – Taille des données de référence avec l'alignement consensuel : le nombre de documents, de phrases et d'occurrences de mots pour chaque sous-corpus et registre, et le taux d'alignement

| | | *Expert* | | | | *Non-expert* | | | | *Taux d'* | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *source* | | *aligné* | | *source* | | *aligné* | | *alignement (%)* | |
| *corpus* | *nb docs* | *nb ph.* | *nb occ.* | *nb paires* | *nb occ.* | *nb ph.* | *nb occ.* | *nb paires* | *nb occ.* | *ex.* | *non-ex.* |
| *Méd.* | 12*2 | 4 416 | 44 709 | 502 | 5 751 | 2 736 | 27 820 | 502 | 10 398 | 18 | 11 |
| *Sci.* | 13*2 | 553 | 8 854 | 112 | 3 166 | 263 | 4 688 | 112 | 3 306 | 20 | 43 |
| *Wiki.* | 14*2 | 2 494 | 36 002 | 49 | 1 100 | 238 | 2 659 | 49 | 853 | 2 | 21 |

# 3 Détection et alignement de phrases parallèles

Les documents sont d'abord segmentés en phrases en exploitant la ponctuation forte (*i.e. . ? ! ; :*). Nous avons aussi retiré, au sein de chaque sous-corpus, les phrases qui apparaissent au moins dans la moitié des documents (typiquement, les mentions légales et les titres de sections) et les phrases sans caractères alphabétiques. Cela réduit la combinatoire de phrases de 940 000 à 590 000 environ.

Nous abordons la détection et l'alignement automatique de phrases parallèles comme un problème de classification, où il s'agit d'assigner chaque paire de phrases analysées à l'une des deux catégories :

— *alignées* : les phrases sont parallèles et peuvent être alignées ;
— *non-alignées* : les phrases ne sont pas parallèles et ne peuvent pas être alignées.

Les données de référence fournissent 663 exemples positifs. Les exemples négatifs contiennent toutes les paires de phrases possibles, pour chaque paire de documents, en excluant les paires de phrases alignées, soit environ 590 000 paires de phrases au total. Nous avons effectué des tests d'alignement de phrases avec plusieurs classifieurs du module `scikit-learn`[2] : *Perceptron* (Rosenblatt, 1958),

2. `https://scikit-learn.org/stable/`

*Multilayer Perceptron* (MLP) (Rosenblatt, 1961), *Linear discriminant analysis* (LDA) (Fisher, 1936), *Quadratic discriminant analysis* (QDA) (Cover, 1965), *Stochastic gradient descent* (SGD), *Linear SVM* (Vapnik & Lerner, 1963). Les modèles obtenus avec le classification binaire issu de la régression logistique montrent les meilleurs résultats. Ce sont donc ces résultats que nous reportons dans la suite de la présentation.

Nos expériences sont basées sur plusieurs types de descripteurs calculés sur les textes non lemmatisés. Plusieurs combinaisons de ces descripteurs ont été testées (Cardon & Grabar, 2018), ce qui nous a permis de sélectionner les descripteurs les plus efficaces :

— Nombre de mots communs, à l'exception des mots vides. Ce descripteur permet de calculer l'intersection lexicale de base entre les versions techniques et simplifiées des phrases ;
— Pourcentage de mots d'une phrase inclus dans l'autre, dans les deux directions. Ce descripteur représente de possibles relations d'inclusion lexicale entre les phrases ;
— Différence de longueur entre les phrases. Ce descripteur vise les inclusions lexicales et suppose que la simplification peut impliquer une association stable avec la longueur des phrases ;
— Différence de la longueur moyenne des mots entre les phrases. Ce descripteur est similaire au précédent mais il prend en compte la différence moyenne de la longueur des phrases ;
— Nombre total de bigrammes et de trigrammes en communs (un descripteur pour chaque). Ce descripteur est calculé sur les n-grammes de caractères. La supposition est que, au niveau de caractères, certaines séquences plus petites que les mots peuvent aussi être partagées par les deux phrases et donc être significatives pour leur alignement ;
— Mesures de similarité (cosinus, Dice et Jaccard). Ce descripteur fournit une indication plus sophistiquée sur l'intersection lexicale des phrases. Le poids de chaque mot est de 1 ;
— Distance d'édition de Levenshtein (Levenshtein, 1966). La distance prend en compte les opérations d'édition de base (insertion, suppression et substitution) au niveau des caractères (acception classique) et au niveau des mots. Le coût de chaque opération est de 1.

Les paires alignées manuellement sont divisées en trois sous-ensembles :

$E$ : 238 paires avec équivalence sémantique,
$TIS$ : 237 paires où le contenu de la phrase technique est entièrement compris dans la phrase simplifiée. La phrase simplifiée contient donc une sémantique supplémentaire,
$SIT$ : 112 paires où le contenu de la phrase simplifiée est entièrement compris dans la phrase technique. La phrase technique contient donc une sémantique supplémentaire.

Les paires de phrases avec d'inclusion ($SIT$ et $TIS$) permettent de repérer les cas lorsque les phrases sont segmentées (une phrase technique est scindée en plusieurs phrases simples) ou fusionnées (plusieurs phrases techniques sont fusionnées en une seule phrase dans la version simplifiée du texte).

Pour chaque sous-ensemble, nous prenons d'abord autant de paires équilibrées que d'exemples négatifs, que nous sélectionnons aléatoirement. Ensuite nous augmentons progressivement le nombre de paires non alignées jusqu'à un ratio de 3000 :1, ratio proche de celui des données réelles. Pour chaque ensemble déséquilibré $D$ ainsi constitué, nous faisons deux expériences :

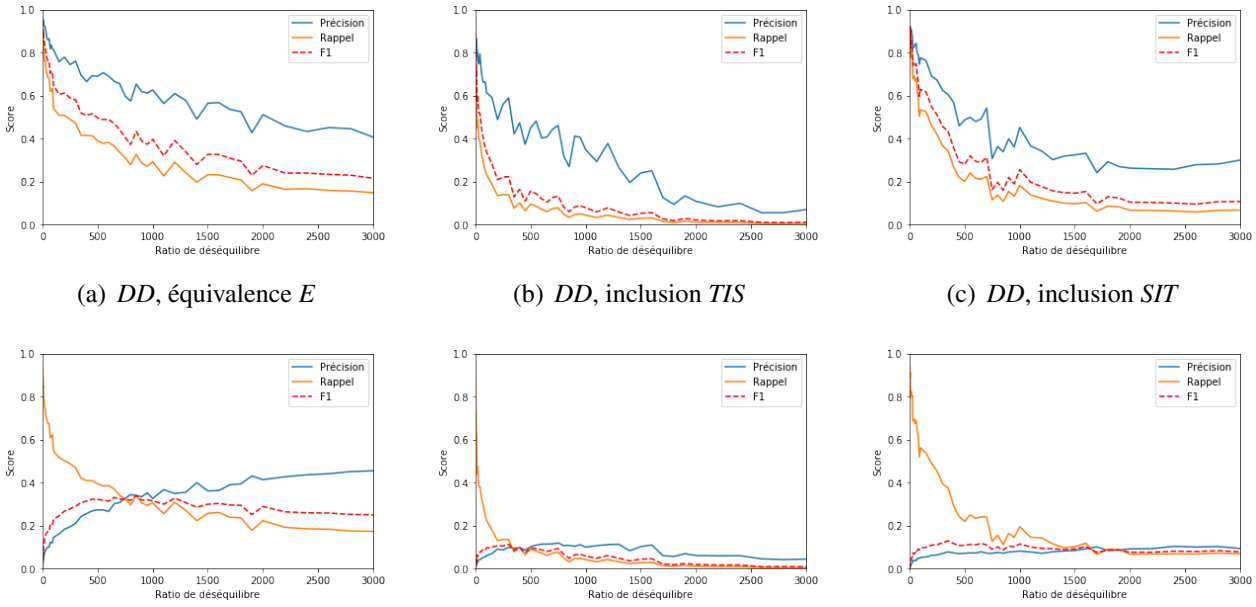
$DD$ : Entraînement et test au sein de l'ensemble déséquilibré $D$ ;
$DR$ : Entraînement sur l'ensemble $D$ et test sur les données réelles $R$. Notons que $D$ est inclus dans $R$.

Pour l'évaluation, nous divisons les données en deux parties : deux tiers pour l'entraînement et un tiers pour le test. L'évaluation est effectuée en calculant le rappel, la précision et la F-mesure. Comme notre objectif vise la détection des paires alignées, les scores sont rapportés uniquement pour la classe des paires alignées. Une autre raison d'exclure les scores de paires non alignées est, qu'avec le déséquilibre qui augmente, cette classe négative obtient très rapidement des scores très elevés (>0,99).

Pour avoir une évaluation plus fiable, chaque expérience est effectuée cinquante fois et les résultats présentés correspondent donc aux valeurs moyennes de ces cinquante itérations. La différence d'une itération à l'autre est due au fait que l'ensemble des paires non alignées est différent à chaque fois car créé aléatoirement.

# 4 Présentation et discussion des résultats

FIGURE 1 – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*).




(a) *DD*, équivalence *E* (b) *DD*, inclusion *TIS* (c) *DD*, inclusion *SIT*

(d) *DR*, équivalence *E* (e) *DR*, inclusion *TIS* (f) *DR*, inclusion *SIT*

Nous présentons les résultats à la figure 1 : l'axe $x$ représente l'augmentation du déséquilibre (seule la première position 1 correspond à des données équilibrées), alors que l'axe $y$ représente les scores de précision, rappel et F-mesure. Les résultats pour les trois sous-ensembles sont présentés : équivalence (figures 1(a) et 1(d)), inclusion *TIS* (figures 1(b) et 1(e)) et inclusion *SIT* (figures 1(c) et 1(f)). La première ligne correspond aux résultats obtenus lorsque l'entraînement et le test sont effectués sur des données avec le même ratio de déséquilibre. La deuxième ligne correspond aux résultats obtenus par les mêmes modèles mais testés sur l'ensemble des données annotées manuellement.

Les paires équivalentes (figures 1(a) et 1(d)) sont plus faciles à catégoriser que les inclusions : d'une part, elles sont assez nombreuses et, d'autre part, elles doivent présenter des modèles de transformation plus stables. Les scores de précision et de rappel sont alors plus élevés à différents points de déséquilibre. Par exemple, au point 1 de la figure 1(a), la F-mesure est de 0,94 (0,96 de précision et 0,93 de rappel). Ce résultat est positif car les phrases équivalentes fournissent les informations les plus utiles et complètes pour décrire les transformations requises lors de la simplification. Avec les relations d'inclusion, au même point, nous avons 0,89 de F-mesure pour *TIS* (0,90 de précision et 0,89 de rappel) et 0,92 de F-mesure pour *SIT* (0,92 de précision et 0,93 de rappel). Nous supposons que des paires d'inclusion couvrent une grande variété de situations, ce qui est également plus difficile à modéliser. Nous prévoyons de faire des filtres supplémentaires pour mieux calibrer les résultats.

Nous voyons donc que les données équilibrées donnent de très bons résultats pour les trois jeux de données (*E, TIS* et *SIT*). En revanche, quand le déséquilibre est introduit, les performances sont réduites. Cela signifie que le déséquilibre crée de la confusion entre les paires alignables et non alignables. Cependant, le déséquilibre a un plus grand impact sur les paires qui relèvent de l'inclusion, que ce soit le sens *TIS* ou *SIT*. Une fois encore, il nous semble que cela est dû au fait que les cas d'inclusion sont beaucoup plus variés que les cas d'équivalence et sont en conséquence plus difficiles à circonscrire. Nos résultats indiquent que, quand on traite des données réelles, il vaut mieux effectuer la classification avec les modèles entraînés sur des données équilibrées. Un autre résultat intéressant est que la précision est plus élevée que le rappel. Cela s'observe particulièrement bien avec les expériences où l'entraînement et le test sont faits avec le même ratio de déséquilibre (figures 1(a), 1(b) et 1(c)). De manière générale, nous voyons que les résultats sont élevés lorsque l'on traite des données équilibrées. Cependant, comme le déséquilibre est une caractéristique naturelle des données que nous traitons, le travail à venir sera concentré sur la mise au point de descripteurs et filtres pour éliminer un maximum de phrases non alignables, *a priori* et/ou *a posteriori* de l'alignement.

# 5 Conclusion et perspectives

Nous proposons des expériences pour la détection et d'alignement de phrases parallèles dans des corpus comparables monolingues en français. L'aspect comparable se situe au niveau de la technicité des documents, qui met en regard des versions techniques et simplifiées des documents traitant du même sujet. Nous utilisons un corpus disponible CLEAR) qui relève du domaine biomédical. Plusieurs expériences sont menées : trois jeux de données (paires équivalentes et inclusion du sens d'une phrase dans l'autre) et données équilibrées et déséquilibrées. Sur les données équilibrées, nous atteignons une F-mesure de 0,94, avec un bon équilibre entre la précision et le rappel. Sur les données déséquilibrées, nous obtenons jusqu'à 0,89 et 0,92 de F-mesure. Les résultats restent meilleurs quand des modèles entraînés sur des données équilibrées sont utilisés. En l'état, cette méthode ne permet pas de générer un corpus de phrases parallèles de façon complètement automatique. Cependant, pour les paires équivalentes, le travail manuel est grandement allégé : nous atteignons un ratio d'environ 40 % de paires alignées correctement dans la sortie de notre classifieur sur de nouvelles données, alors que ce ratio est d'environ 0,025 % dans les données brutes (environ 4 000 paires non alignées pour 1 paire alignée). À l'avenir, nous prévoyons d'exploiter les meilleurs modèles générés pour enrichir le corpus de phrases parallèles. Les scores de rappel peuvent correspondre aux mesures de référence pour choisir le meilleur classifieur. Une attention particulière sera apportée au filtrage des données *a priori* et/ou *a posteriori* de l'alignement. D'autres pistes de travail concernent l'exploitation d'autres descripteurs et méthodes pour l'alignement de phrases. Comme la description du corpus le montre, la distance lexicale entre les phrases techniques et simplifiées est assez élevée. En conséquence, nous comptons nous tourner vers l'utilisation de plongements lexicaux ainsi que vers l'exploitation de connaissances externes pour pallier cette difficulté.

# Remerciements

La présente publication s'inscrit dans le projet *CLEAR (Communication, Literacy, Education, Accessibility, Readability)* financé par l'ANR sous la référence ANR-17-CE19-0016-01. Nous remercions les relecteurs pour leurs remarques constructives.

# Références

ABDUL-RAUF S. & SCHWENK H. (2009). On the use of comparable corpora to improve SMT performance. In *European Chapter of the ACL*, p. 16–23.

AGIRRE E., BANEA C., CARDIE C., CER D., DIAB M., GONZALEZ-AGIRRE A., GUO W., LOPEZ-GAZPIO I., MARITXALAR M., MIHALCEA R., RIGAU G., URIA L. & WIEBE J. (2015). SemEval-2015 task 2 : Semantic textual similarity, english, spanish and pilot on interpretability. In *SemEval 2015*, p. 252–263.

AGIRRE E., BANEA C., CER D., DIAB M., GONZALEZ-AGIRRE A., MIHALCEA R., RIGAU G. & WIEBE J. (2016). SemEval-2016 task 1 : Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval 2016*, p. 497–511.

AGIRRE E., CER D., DIAB M., GONZALEZ-AGIRRE A. & GUO W. (2013). *sem 2013 shared task : Semantic textual similarity. In **SEM*, p. 32–43.

AMA (1999). Health literacy : report of the council on scientific affairs. Ad hoc committee on health literacy for the council on scientific affairs, American Medical Association. *JAMA*, **281**(6), 552–7.

BARZILAY R. & ELHADAD N. (2003). Sentence alignment for monolingual comparable corpora. In *EMNLP*, p. 25–32.

CARDON R. & GRABAR N. (2018). Identification of parallel sentences in comparable monolingual corpora from different registers. In *LOUHI 2018*, p. 1–11.

CHEN P., ROCHFORD J., KENNEDY D. N., DJAMASBI S., FAY P. & SCOTT W. (2016). Automatic text simplification for people with intellectual disabilities. In *AIST*, p. 1–9.

CLOUGH P., GAIZAUSKAS R., PIAO S. S. & WILKS Y. (2002). METER : Measuring text reuse. In *ACL*, p. 152–159.

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1), 37–46.

COVER T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, **14**(3), 326–334.

FERNANDO S. & STEVENSON M. (2008). A semantic similarity approach to paraphrase detection. In *Comp Ling UK*, p. 1–7.

FISHER R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**(2), 179–188.

GANITKEVITCH J., VAN DURME B. & CALLISON-BURCH C. (2013). PPDB : The paraphrase database. In *NAACL-HLT*, p. 758–764.

GRABAR N. & CARDON R. (2018). CLEAR – Simple Corpus for Medical French. In *Workshop on Automatic Text Adaption (ATA)*, p. 1–11.

GUO W. & DIAB M. (2012). Modeling sentences in the latent space. In *ACL*, p. 864–872.

HE H., GIMPEL K. & LIN J. (2015). Multi-perspective sentence similarity modeling with convolutional neural networks. In *EMNLP*, p. 1576–1586, Lisbon, Portugal.

KIROS R., ZHU Y., SALAKHUTDINOV R., ZEMEL R. S., TORRALBA A., URTASUN R. & FIDLER S. (2015). Skip-thought vectors. In *Neural Information Processing Systems (NIPS)*, p. 3294–3302.

KOEHN P. (2005). Europarl : A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings : the tenth Machine Translation Summit*, p. 79–86, Phuket, Thailand : AAMT AAMT.

KUMANO T., TANAKA H. & TOKUNAGA T. (2007). Extracting phrasal alignments from comparable corpora by using joint probability SMT model. In *Int Conf on Theoretical and Methodological Issues in Machine Translation*.

LAI A. & HOCKENMAIER J. (2014). Illinois-LH : A denotational and distributional approach to semantics. In *Workshop on Semantic Evaluation (SemEval 2014)*, p. 239–334, Dublin, Ireland.

LEROY G., KAUCHAK D. & MOURADI O. (2013). A user-study measuring the effects of lexical simplification and coherence enhancement on perceived and actual text difficulty. *Int J Med Inform*, **82**(8), 717–730.

LEVENSHTEIN V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet physics. Doklady*, **707**(10).

MADNANI N., TETREAULT J. & CHODOROW M. (2012). Re-examining machine translation metrics for paraphrase identification. In *NAACL-HLT*, p. 182–190.

MCGRAY A. (2005). Promoting health literacy. *J of Am Med Infor Ass*, **12**, 152–163.

MIHALCEA R., CORLEY C. & STRAPPARAVA C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, p. 1–6.

MILLER G. A., BECKWITH R., FELLBAUM C., GROSS D. & MILLER K. (1993). *Introduction to WordNet : An On-line Lexical Database*. Rapport interne, WordNet.

MUELLER J. & THYAGARAJAN A. (2016). Siamese recurrent architectures for learning sentence similarity. In *AAAI Conference on Artificial Intelligence*, p. 2786–2792.

MUNTEANU D. S. & MARCU D. (2002). Processing comparable corpora with bilingual suffix trees. In *EMNLP*, p. 289–295.

MUNTEANU D. S. & MARCU D. (2005). Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, **31**(4), 477–504.

MUNTEANU D. S. & MARCU D. (2006). Extracting parallel sub-sentential fragments from non-parallel corpora. In *COLING-ACL*, p. 81–88.

NELKEN R. & SHIEBER S. M. (2006). Towards robust context-sensitive sentence alignment for monolingual corpora. In *EACL*, p. 161–168.

PAETZOLD G. H. & SPECIA L. (2016). Benchmarking lexical simplification systems. In *LREC*, p. 3074–3080.

QIU L., KAN M.-Y. & CHUA T.-S. (2006). Paraphrase recognition via dissimilarity significance classification. In *Empirical Methods in Natural Language Processing*, p. 18–26, Sydney, Australia.

ROSENBLATT F. (1958). The perceptron : a probabilistic model for information storage and organization in the brain. *Psychological Review*, **65**(6), 386–408.

ROSENBLATT F. (1961). *Principles of Neurodynamics : Perceptrons and the Theory of Brain Mechanisms*. Washington DC : Spartan Books.

RUDD E. (2013). Needed action in health literacy. *J Health Psychol*, **18**(8), 1004–10.

SEVERYN A., NICOSIA M. & MOSCHITTI A. (2013). Learning semantic textual similarity with structural representations. In *Annual Meeting of the Association for Computational Linguistics*, p. 714–718.

TAI K. S., SOCHER R. & MANNING C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. In *Annual Meeting of the Association for Computational Linguistics*, p. 1556–1566, Beijing, China.

TSUBAKI M., DUH K., SHIMBO M. & MATSUMOTO Y. (2016). Non-linear similarity learning for compositionality. In *AAAI Conference on Artificial Intelligence*, p. 2828–2834.

UTIYAMA M. & ISAHARA H. (2003). Reliable measures for aligning Japanese-English news articles and sentences. In *Annual Meeting of the Association for Computational Linguistics*, p. 72–79.

VAPNIK V. & LERNER A. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, **24**, 709–715.

VU T. T., TRAN G. B. & PHAM S. B. (2014). Learning to simplify children stories with limited data. In L. . SPRINGER, Ed., *Intelligent Information and Database Systems*, p. 31–41.

WAN S., DRAS M., DALE R. & PARIS C. (2006). Using dependency-based features to take the ”para-farce” out of paraphrase. In *Australasian Language Technology Workshop*, p. 131–138.

YANG C. C. & LI K. W. (2003). Automatic construction of English/Chinese parallel corpora. *J. Am. Soc. Inf. Sci. Technol.*, **54**(8), 730–742.

ZHANG Y. & PATRICK J. (2005). Paraphrase identification by text canonicalization. In *Australasian Language Technology Workshop*, p. 160–166.

ZHAO J., ZHU T. T. & LAN M. (2014). ECNU : One stone two birds : Ensemble of heterogenous measures for semantic relatedness and textual entailment. In *Workshop on Semantic Evaluation (SemEval 2014)*, p. 271–277.

ZHU Z., BERNHARD D. & GUREVYCH I. (2010). A monolingual tree-based translation model for sentence simplification. In *COLING 2010*, p. 1353–1361.

# Développement d'un lexique morphologique et syntaxique de l'ancien français

Benoît Sagot
Inria, 2 rue Simone Iff, 75012 Paris, France
benoit.sagot@inria.fr

RÉSUMÉ

Nous décrivons dans cet article notre travail de développement d'un lexique morphologique et syntaxique à grande échelle de l'ancien français pour le traitement automatique des langues. Nous nous sommes appuyés sur des ressources dictionnairiques et lexicales dans lesquelles l'extraction d'informations structurées et exploitables a nécessité des développements spécifiques. De plus, la mise en correspondance d'informations provenant de ces différentes sources a soulevé des difficultés. Nous donnons quelques indications quantitatives sur le lexique obtenu, et discutons de sa fiabilité dans sa version actuelle et des perspectives d'amélioration permises par l'existence d'une première version, notamment au travers de l'analyse automatique de données textuelles.

ABSTRACT

**Development of a morphological and syntactic lexicon of Old French.**

In this paper we describe our work on the development of a large-scale morphological and syntactic lexicon of Old French for natural language processing. We rely on dictionary and lexical resources, from which the extraction of structured and exploitable information required specific developments. In addition, matching information from these different sources posed difficulties. We provide quantitative information on the resulting lexicon, and discuss its reliability in its current version and the prospects for improvement allowed by the existence of a first version, in particular through the automatic analysis of textual data.

MOTS-CLÉS : Lexique morphologique, Lexique syntaxique, Ancien français.

KEYWORDS: Morphological lexicon, Syntactic lexicon, Old French.

# 1 Introduction

L'ancien français regroupe l'ensemble des variétés romanes qualifiées de langues d'oïl, qui se sont développées au nord de la France, au sud de la Belgique et dans les îles Anglo-Normandes [1], telles qu'elles étaient parlées du VIIIe au milieu du XIVe siècle environ. L'ancien français se distingue notamment du moyen français, qui lui fait suite, par la présence de déclinaisons nominales. Ancien puis moyen français peuvent être vus comme les ancêtres successifs du français contemporain.

Les deux principales bases de données textuelles, étiquetées semi-automatiquement en parties du discours et en lemmes, sont la Base de Français Médiéval, ci-après BFM (Guillot *et al.*, 2017) [2],

1. Et jusqu'en Angleterre, si l'on tient compte par exemple des lais de Marie de France.
2. http://bfm.ens-lyon.fr

et le Nouveau Corpus d'Amsterdam, ci-après NCA (Stein & al., 2008). Ces bases contiennent respectivement plus de 4 millions et plus de 3 millions de mots. Les deux principaux corpus arborés de l'ancien français sont le Syntactic Reference Corpus of Medieval French, ou SRCMF (Stein & Prévost, 2013) et la partie couvrant l'ancien français au sein du corpus MCVF (Martineau, 2008). Ces deux corpus, dont seul le premier est librement téléchargeable, ne sont pas annotés selon le même guide d'annotation. Enfin, une partie du SRCMF a fait l'objet d'un travail de conversion vers le modèle de l'initiative *Universal Dependencies* (UD) [3]. Toutes ces ressources rassemblent des textes variés, tant sur le plan du style (prose, vers), du genre (littéraire, religieux, historique, didactique), de l'époque (du Xe au XIIIe siècle) que de la géographie dialectale. Toutefois, certains biais sont inévitables, qui affectent en particulier les études linguistiques quantitatives. Ainsi, faute de textes en prose, les premiers siècles de la période ne peuvent être couverts que par des textes en vers.

La disponibilité de ces corpus a permis le développement d'études linguistiques quantitatives et d'outils de TAL. Des expériences d'annotation morphosyntaxiques ont notamment été réalisées par Stein (2014) avec TreeTagger, suivies par celles de Guibon *et al.* (2014, 2015) avec des champs aléatoires conditionnels. Ces expériences ont été toutes deux complétées par une annotation syntaxique en dépendances à l'aide de l'analyseur Mate (Bohnet, 2010) entraîné sur le SRCMF.

Toutefois, le développement de ressources lexicales pour l'ancien français destinées au traitement automatique des langues n'est pas aussi avancé. Pour l'ancien français, on ne peut guère citer que FROLEX [4], lexique librement disponible développé dans le cadre du projet PaLaFra et que ses auteurs définissent comme un lexique morphologique du français du IXe au XVe siècle — mais nous reviendrons sur l'applicabilité en l'espèce de la notion de « lexique morphologique ». Il a été construit automatiquement à partir de textes annotés extraits de la BFM et du NCA, mais également à partir du Dictionnaire de Moyen Français (DMF) [5, 6]. D'autres ressources existent, de natures différentes, notamment dictionnairiques. Nous y reviendrons à la prochaine section. Aucune de ces ressources ne constitue véritablement un lexique morphologique, pas plus qu'un lexique syntaxique. Un tel lexique est pourtant indispensable au développement de certains analyseurs syntaxique. Il permettrait l'amélioration d'outils comme les étiqueteurs morphosyntaxiques, et ouvrirait de nouvelles perspectives en linguistique quantitative, y compris diachronique.

C'est au développement d'un tel lexique que cet article est consacré. Ce lexique, nommé OFrLex, est donc un lexique morphologique et syntaxique de l'ancien français [7], complété par des liens vers les ressources de départ et des gloses pour certaines entrées. Les difficultés principales ont été de trois ordres : (1) la difficulté de transformer des ressources faiblement structurées en données structurées, (2) la non-cohérence des façons dont sont représentées les informations lexicales, rendant délicate la mise en correspondances entre entrées traitant d'un même lexème, et (3) la construction d'informations précises (classes morphologiques, cadres de sous-catégorisation) à partir de ressources ne contenant ces informations que de façon partielle et sous-spécifiée, voire ne les contenant pas du tout. Il en a résulté le développement d'heuristiques et d'outils dédiés et un effort manuel important. Cela fait d'OFrLex un lexique développé ni de façon automatique ni de façon entièrement manuelle.

3. Ce travail de conversion et leurs auteurs sont détaillés sur le site GitHub du corpus converti.

4. `https://github.com/sheiden/Medieval-French-Language-Toolkit`

5. `http://www.atilf.fr/dmf/`

6. Les lexiques LGeRM, développés à l'ATILF, existent sous deux variantes, l'une dite « médiévale » couvre la période 1300-1550, l'autre couvre les XVIe et XVIIe siècles. Ils ne couvrent donc pas l'ancien français.

7. Comme nous le verrons, le volet syntaxique est dans ce travail restreint au lexique verbal. La valence nominale et adjectivale, en particulier, n'est pas encore décrite. C'est naturellement une première étape, mais elle est à la fois nécessaire et, dans un premier temps, suffisante pour le développement d'analyseurs syntaxiques s'appuyant sur des grammaires lexicalisées.

| Forme | Fréquence | | Étiquette d'origine | | | Étiquette CATTEXT étendue | | Lemme | Source |
|---|---|---|---|---|---|---|---|---|---|
| | BFM | DMF | AFRLEX | BFM | DMF | conv. 1 | conv. 2 | | du lemme |
| abassera | 2 | 0 | | *no pos* | | *no pos* | OUT | *no lemma* | BFM |
| abasseur | 0 | 0 | NOM | | subst. masc. | NOMcom | NOMcom | abasseur | DMF |
| abasseure | 0 | 0 | | | verbe | | VER | abasseurer | DMF |
| gaiement | 0 | 9 | | | adv. | | ADV | gaiement | DMF |
| gaiement | 1 | 0 | | ADVgen | | ADVgen | APD | *no lemma* | BFM |

TABLE 1 – Quelques exemples d'entrées de FROLEX

Nous présentons tout d'abord les ressources dont nous sommes partis pour développer ce lexique et la façon dont nous en avons extrait des entrées lexicales structurées. Nous décrivons ensuite comment nous en avons dérivé OFrLex. Nous terminons par des données quantitatives sur OFrLex suivies d'une discussion sur sa fiabilité et les étapes futures de son développement et de son utilisation.

# 2 Extraction d'informations à partir de sources hétérogènes

Comme indiqué ci-dessus, **FROLEX** est une compilation de ressources provenant de la BFM, du NCA et du DMF. Il est constitué de plus d'un million d'entrées extensionnelles, c'est-à-dire d'entrées correspondant chacune à une graphie particulière d'une forme fléchie donnée. Les informations disponibles pour chaque entrée varient d'une entrée à l'autre, en fonction de sa source. Les parties du discours ont toutes été converties, parfois de façon sous-spécifiée, dans le modèle CATTEX, étendu par des indications de genre et de nombre, lorsque pertinent. Quelques exemples d'entrées sont montrées dans la table 1, où l'on constate que certaines entrées sont bruitées. De plus, la variété des sources induit des incohérences dans les conventions de lemmatisation, lorsque toutefois un lemme est fourni. Enfin, l'utilisation du DMF comme source a pour conséquence que ce lexique mélange des entrées relevant de l'ancien français et des entrées relevant du moyen français. Il ne s'agit donc pas d'un lexique morphologique à proprement parler : les lemmes présents dans la ressource n'y sont pas représentés par toutes leurs formes fléchies, et certaines formes ne sont pas associées à des informations morphologiques au-delà de la seule étiquette CATTEX.

Pour l'ancien français, le **Wiktionary (anglophone)** contient environ 6500 entrées lexicales intentionnelles (une entrée correspond à un lexème), ainsi que des descriptions formalisées de classes flexionnelles. Par exemple, l'entrée pour *mengier*[8] fournit des formes alternatives (comme *mangier*), une étymologie, une glose en anglais, et les informations nécessaires pour définir sa flexion. Le processus d'extraction que nous avons utilisé est inspiré par celui décrit dans (Sagot, 2014). Nous avons tout d'abord converti le Wiktionary brut, au format wiki, en un fichier XML structuré. Nous avons ensuite extrait des entrées morphologiques complètes à partir de ce fichier XML : chaque entrée est constituée d'une forme de citation, d'un identifiant de classe flexionnelle et de la liste des radicaux ou formes irréguliers le cas échéant. Nous avons alors développé manuellement dans le formalisme Alexina PYFRBLI (Sagot & Walther, 2013) une grammaire morphologique qui décrit formellement les plus importantes des classes flexionnelles utilisées par Wiktionary.

Le ***Altfranzösisches Wörterbuch*** **de Tobler et Lommatzsch, ci-après TL**, est le dictionnaire d'ancien français de référence. Ses articles sont rédigés en allemand. Nous l'avons utilisé sous deux

8. https://en.wiktionary.org/wiki/mengier#Old_French

| Lemma<br>*Lemme* | Haupt-eintrag<br>*Entrée princ.* | Wortart<br>*Catégorie* | Var.<br>*Var.* | Werk<br>*Source* | Band<br>*Volume* | Spalte<br>*Page* | Zeile<br>*Ligne* | IstVar.<br>*Est une var.* |
|---|---|---|---|---|---|---|---|---|
| aatir | | v. | ahatir | tl | 1 | 31 | 37 | 0 |
| aatir | aaatir | v. | | | 1 | 25 | 32 | 1 |
| aatir | atir | v. | | | 1 | 640 | 52 | 1 |
| aatise | | s.f. | | tl | 1 | 33 | 34 | 0 |
| aatison | | s.f. | | tl | 1 | 33 | 37 | 0 |

TABLE 2 – Quelques exemples partiels extraits de l'index des entrées du Tobler-Lommatzsch

| | |
|---|---|
| ealemlne s. f. s. chalemine.<br>calemon $. m.<br>[Name eines Vogels : s. A. Delboulle, Rom. XXXI 366 ; A. Thomas, eb. XXXVI 25 260.]<br>calende s. /. s. chalende.<br>calendre s. /. s. chalendre.<br>ealer (nfr. caler) vb.<br>[REW 1487 cafare ; Godefroy VIII 30 (Compl.) 412a]<br>trans. (Segel) niederlassen, streichen : Therfés s'escrie : Cale, cale ! Mes tuit li | calemine s. f., s. chalemine.<br>calemon s. m.<br>[Name eines Vogels : s. A. Delboulle, Rom. XXXI 366 ; A. Thomas, eb. XXXVI 25 260.]<br>calende s. f., s. chalende.<br>calendre s. f., s. chalendre.<br>caler (nfz. caler) vb.<br>[REW 1487 cafare ; Godefroy VIII (Compl.) 412a]<br>trans. (Segel) niederlassen, streichen : Therfés s'escrie : Cale, cale ! Mes tuit li |

| | | | |
|---|---|---|---|
| calemine | NOMcom.f | s.f. | |
| calemon | NOMcom.m | s.m. | Name eines Vogels |
| calende | NOMcom.f | s.f. | |
| calendre | NOMcom.f | s.f. | |
| caler | VER | vb. | [trans.] (Segel) niederlassen, streichen |

TABLE 3 – Extrait du Tobler-Lommatzsch OCRisé avant (gauche) et après (droite) correction partielle, et entrées structurées extraites (bas)

formes, toutes deux produites et distribuées par Achim Stein [9] :

— Une liste des lemmes, constituée en partie manuellement et dans laquelle nous avons trouvé très peu d'erreurs, complétée par un index des formes du *Dictionnaire* de Godefroy pour la fin de l'alphabet [10] (la colonne « Werk » indique la source de chaque ligne : « tl » pour le TL et « g » pour le Dictionnaire de Godefroy). Quelques entrées (simplifiées) de cet index sont fournis à la table 2. On notera que la liste des lemmes distingue les entrées principales (« Haupteintrag ») et les entrées secondaires, ou variantes (généralement des variantes graphiques) : toute entrée secondaire est associée à son entrée principale, et les références (page et ligne) sont données pour l'entrée principale et pour l'entrée secondaire.

— Une version complète OCRisée, qui contient un très grand nombre d'erreurs d'OCR. Nous avons donc corrigé partiellement cette version, de façon manuelle mais systématique, en mettant l'accent sur les parties cruciales des entrées, telles que le type de mot (cf. table 3). Ce travail a été réalisé en alternance avec le développement d'un extracteur d'entrées structurées à partir de la version corrigée du TL OCRisée : cet extracteur effectuant de nombreuses vérifications formelles, il refuse de traiter une entrée dans laquelle il identifie des erreurs. Le résultat de l'extraction, qui inclut des catégories CATTEX, est illustré en bas de la table 3.

Nous avons également utilisé le ***Lexique de l'ancien français*** **de Godefroy**, dans une version publiée sur Wikisource [11], construite au moyen d'un OCR de très bonne qualité et partiellement corrigée

9. https://www.ling.uni-stuttgart.de/institut/ilr/toblerlommatzsch/downloads.htm
10. Ce dictionnaire n'est pas la même ressource que le *Lexique* décrit plus bas.
11. https://fr.wikisource.org/wiki/Lexique_de_l'ancien_français

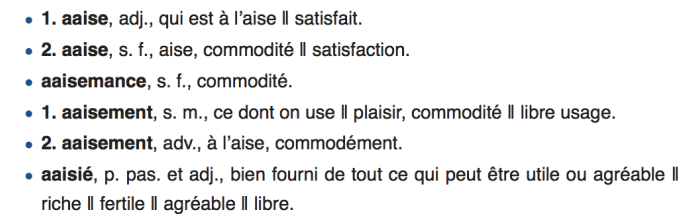



FIGURE 1 – Extrait du *Lexique* de Godefroy dans sa version Wikisource

(cf. table 1). Cette ressource est très couvrante mais connue pour contenir des mots (et sens) fantômes. Nous avons donc filtré au moyen de la *Base des mots fantômes [du Godefroy]* [12]. De plus, il couvre jusqu'au XVe siècle, débordant ainsi sur le moyen français. Le caractère structurée du *Lexique* nous a permis d'extraire facilement des entrées structurées combinant une forme de citation, une catégorie CATTEX (complétée le cas échéant d'une information de genre), une définition et l'indication du volume et de la page correspondante.

Enfin, le **Dictionnaire Électronique de Chrétien de Troyes (DÉCT)** est un dictionnaire complet de cet écrivain du XIIe siècle distribué par le CNRTL au format PDF. Nous l'avons converti en format texte, et en avons extrait de façon semi-automatique des entrées structurées à l'aide de règles simples. L'un des intérêts du DÉCT est qu'il relie explicitement ses entrées à des entrées d'autres dictionnaires, dont le TL et le *Dictionnaire* de Godefroy (dont les lemmes sont plus ou moins les mêmes que dans son *Lexique*). Il fournit également les graphies des formes fléchies attestées de chaque entrée.

# 3 Combinaison des sources et création d'OFrLex

Nous avons tout d'abord relié entre elles les entrées structurées extraites des différentes ressources de la façon suivante, en utilisant comme formes de citation celles du TL. Les entrées du DÉCT pointent vers des entrées du TL, avec très peu d'erreurs. Pour les autres ressources, les lemmes ayant plusieurs entrées conduisent à de multiples correspondances possibles entre entrées issues du Godefroy, du TL et du DÉCT. De plus, pour certains lemmes qui n'ont qu'un lemme dans chaque ressource, ces entrées ne correspondent pas : il faut créer plusieurs entrées et non pas les fusionner. Nous avons donc procédé à une désambiguïsation manuelle, en nous appuyant pour cela sur les définitions contenues dans les différentes ressources. Les informations morphologiques sont extraites des entrées issues du Wiktionary, ou rajoutées automatiquement (lorsque la forme de citation rend cela possible) ou manuellement. Des variantes de formes sont associées à l'entrée grâce aux informations extraites de FROLEX. Le résultat de ce travail est un lexique morphologique où une entrée, qui correspond à un lexème (et non seulement à un lemme) est reliée aux entrées dans les ressources de départ et complétée par les informations que nous y avons extraites, notamment des définitions et gloses, ainsi que des informations sur les variantes (une entrée de type « variante » extraite du TL est associée à son entrée principale, laquelle liste ses variantes). Une catégorie UD, ou UPOS, est également ajoutée automatiquement sur la base des catégories extraites à partir des ressources de départ.

Nous avons alors cherché à compléter les entrées verbales de ce lexique par une couche syntaxique, en suivant les mêmes conventions et critères que ceux du lexique Alexina du français contemporain, le Le*fff* (Sagot, 2010). Pour cela, nous avons attribué à chaque entrée verbale les informations syntaxiques (valence…) d'un verbe du Le*fff* susceptible d'être syntaxiquement similaire, ce qui

12. `http://stella.atilf.fr/scripts/fantomes.exe`

**afiner$_1$** v-er 100;Lemma;v;<Suj:cln|sn>;upos=VERB,cat=v;%actif
# <link src="TL" loc="TL:1:189:5+1:1224:51" entry="afiner1" ms="v." def="[intr.] enden || [mit pers. obj.] jem. den Garaus machen || [trans. mit sâchl obj.] beenden, zu Ende führen"/> <syntinfosource via="tldef" synttype="I"/>

**afiner$_2$** v-er 100;Lemma;v;<Suj:cln|sn,Obj:(cla|sn)>;upos=VERB,cat=v;%actif,%passif
# <link src="TL" loc="TL:1:189:47+1:1224:52" entry="afiner2" ms="v." def="[trans.] läutern"/>
<syntinfosource via="tldef" synttype="T"/><hasvariant lemma="effiner" id="1" cat="VER"/>

**effiner** v-er 100;Lemma;v;<Suj:cln|sn,Obj:(cla|sn)>;upos=VERB,cat=v;%actif,%passif
# <link src="TL" loc="TL:1:189:47" entry="afiner2" ms="v." def="[trans.] läutern"/>
<syntinfosource via="tldef" synttype="T"/><variantof lemma="afiner" id="2" cat="VER"/>

**effiner**, verbe du premier groupe à la flexion régulière
Verbe transitif passivable à sujet nominal ou clitique et à objet direct facultatif nominal ou clitique
Variante de afiner$_2$
Entrée correspondante dans le Tobler-Lommatzsch: afiner2 (1:189:47) '[trans.] läutern'
La valence transitive a été inférée à partir de la glose ci-dessus fournie par le Tobler-Lommatzsch

TABLE 4 – Exemples d'entrées dans OFrLex, suivies d'une version plus explicite de la dernière entrée

| UPOS | ADJ | ADP | ADV | CCONJ | DET | INTJ | NOUN | PRON | PROPN | PUNCT | SCONJ | VERB |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| #entrées | 7895 | 286 | 1848 | 37 | 296 | 205 | 44084 | 517 | 1948 | 19 | 53 | 16817 |
| #lemmes dist. | 7740 | 283 | 1804 | 37 | 259 | 174 | 41191 | 411 | 1934 | 17 | 50 | 16152 |

TABLE 5 – Informations quantitatives sur les entrées d'OFrLex

suppose de lier chaque verbe d'OFrLex à un verbe du Le*fff*. Pour cela, pour chaque entrée verbale d'OFrLex, nous avons utilisé, par ordre de priorité décroissant, l'une des informations suivantes :

— Une « pseudo-glose » rajoutée manuellement, choisie exprès pour ses propriétés syntaxiques supposées identiques (ou similaires) à celles de l'entrée d'OFrLex ;
— Une glose en français contemporain issue d'une des ressources de départ ou ajoutée à la main ;
— La définition issue du Godefroy ou du DÉCT dès lors qu'elle est formée d'un seul mot ;
— Un descendant en français contemporain, extrait du Wiktionary ou rajouté manuellement ;
— Une entrée du Le*fff* dont la forme de citation est identique à celle d'OFrLex.

Si aucune de ces stratégies n'est applicable, nous avons utilisé les indications de valence extraites des définitions du Godefroy, du TL ou du DÉCT, qui contiennent souvent des étiquettes tels que « trans. », « I » (pour « *intransitiv* » dans le TL) ou « refl. » (sous de multiples variantes). En l'absence de telles informations, nous avons considéré par défaut l'entrée comme transitive simple. On notera que les relations entre variantes sont prises en compte pour récupérer la « meilleure » information possible.

La table 4 contient trois entrées verbales, dont la troisième est une variante de la seconde. Elles sont en deux parties : tout d'abord des entrées Alexina classiques (forme de citation, classe flexionnelle, informations syntaxiques), puis, dans la partie « commentaires » de l'entrée (après le signe #), des éléments XML encodant les informations complémentaires. Pour interpréter les informations syntaxiques on pourra se référer à (Sagot, 2010).

# 4 Éléments d'évaluation

Des informations quantitatives sur les entrées d'OFrLex, qui correspondent à des lexèmes, sont fournies à la table 5 par catégorie UD. Alexina permet de produire automatiquement à partir de ces entrées intentionnelles près d'un million entrées extensionnelles décrivant chaque (variante de) chaque forme fléchie de chaque lexème.

Nous avons évalué l'impact de l'utilisation d'OFrLex par un étiqueteur en parties du discours. Une telle évaluation est naturellement très limitée, ne serait-ce que parce qu'elle ne fait pas usage du niveau syntaxique du lexique. Nous avons utilisé à cette fin alVWTagger [13], que nous avons développé dans le cadre de notre participation à la campagne d'évaluation CoNLL 2017 dédiée à l'analyse syntaxique multilingue (Villemonte de La Clergerie *et al.*, 2017). Il s'agit d'un étiqueteur statistique qui, comme son prédécesseur MElt (Denis & Sagot, 2012), peut s'appuyer sur un lexique externe pour produire des traits qui viennent en complément des traits extraits du corpus d'apprentissage ou de test. Nous avons utilisé comme données pour l'apprentissage la section d'entraînement de la version *Universal Dependencies* (v2.4, Nivre & *al.*, 2019) du SRCMF, et la section de développement de ce même corpus comme données d'évaluation. Nous avons entraîné avec alVWTagger deux modèles d'étiquetage en catégories UD, l'un sans utilisation d'OFrLex et l'autre en utilisant un lexique de formes fléchies extrait d'OFrLex dans lequel chaque forme est associée à sa partie du discours OFrLex. Les résultats sont probants : l'utilisation d'OFrLex fait passer l'exactitude globale de 93,8% à 94,8%, et l'exactitude sur les seuls mots inconnus du corpus d'entraînement, qui représentent 8,5% des 16 463 mots du corpus d'évaluation, de 81,6% à 85,7%.

# 5 Discussion

Dans sa version actuelle (version 1), OFrLex n'est pas encore une ressource fiable, malgré sa large couverture. Au niveau morphologique, les classes flexionnelles fournies sont fiables, mais certains verbes irréguliers sont encore imparfaitement décrits. Les informations sémantiques (gloses, définitions, liens vers les entrées des ressources de départ) sont assez fiables, mais certains exemples à revoir ont déjà été identifiés. Enfin, au niveau syntaxique, l'approche décrite ci-dessus est rudimentaire : elle n'a pour vocation que de produire une première version du lexique, version dont l'existence permet d'utiliser des techniques d'amélioration par l'utilisation d'OFrLex dans un analyseur syntaxique.

En effet, le développement d'OFrLex se fait dans un contexte plus large, et va de pair avec celui d'un analyseur syntaxique à large couverture qui s'appuie sur les informations lexicales morphologiques et syntaxiques qu'il fournit. Il s'agit d'un analyseur hybride inspiré de l'analyseur FRMG du français contemporain (Villemonte de la Clergerie, 2005) et dont les premières étapes du développement sont décrites par (Regnault, 2019). Comme FRMG, cet analyseur s'appuie sur une métagrammaire développée à la main, compilée automatiquement en une grammaire d'adjonction d'arbres (TAG) factorisée et avec contraintes. À partir de cette grammaire est produit un analyseur syntaxique symbolique ambigü, complété par un mécanisme de désambiguïsation heuristique ou appris automatiquement à partir d'un corpus arboré. L'analyse automatique de textes par cet analyseur syntaxique pour l'ancien français permettra d'identifier des erreurs et des manques dans OFrLex, soit en comparant des analyses produites avec des analyses de référence lorsque c'est possible, soit par exemple au moyen de techniques de fouille d'erreurs telles que proposées par Sagot & Villemonte de La Clergerie (2008).

# Remerciements

Ce travail a été réalisé dans le cadre du projet ANR-16-CE38-0010 PROFITEROLE (2017–2020) dirigé par Sophie Prévost.

13. `https://gitlab.inria.fr/almanach/alTextProcessing/alAnalyser`

# Références

BOHNET B. (2010). Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, p. 89–97, Beijing, Chine.

DENIS P. & SAGOT B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. *Language Resources and Evaluation*, **46**(4), 721–736.

GUIBON G., TELLIER I., CONSTANT M., PRÉVOST S. & GERDES K. (2014). Parsing Poorly Standardized Language Dependency on Old French. In *Proceedings of the 13th International Workshop on Treebanks and Linguistic Theories (TLT13)*, p. 51–61, Tübingen, Allemagne.

GUIBON G., TELLIER I., PRÉVOST S., CONSTANT M. & GERDES K. (2015). Analyse syntaxique de l'ancien français : quelles propriétés de la langue influent le plus sur la qualité de l'apprentissage ? In *Actes de la 22ème conférence sur le Traitement Automatique du Langage Naturel (TALN)*, Caen, France.

GUILLOT C., HEIDEN S. & LAVRENTIEV A. (2017). Base de français médiéval : une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique. *Diachroniques. Revue de Linguistique française diachronique*, **7**, 168–184.

MARTINEAU F. (2008). Un corpus pour l'analyse de la variation et du changement linguistique. *Corpus*, **7**.

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

REGNAULT M. (2019). Adapting a Metagrammar for Contemporary French to Medieval French. In *TALN-RECITAL 2019 - 26e édition de la conférence TALN (Traitement Automatique des Langues Naturelles) et 21e édition de la conférence jeunes chercheur×euse×s RECITAL*, Toulouse, France.

SAGOT B. (2010). The Le*fff*, a freely available and large-coverage morphological and syntactic lexicon for French. In *7th international conference on Language Resources and Evaluation (LREC 2010)*, La Valette, Malte.

SAGOT B. (2014). DeLex, a freely-avaible, large-scale and linguistically grounded morphological lexicon for German. In *Language Resources and Evaluation Conference*, Reykjavik, Islande : European Language Resources Association.

SAGOT B. & VILLEMONTE DE LA CLERGERIE É. (2008). Fouille d'erreurs sur des sorties d'analyseurs syntaxiques. *Traitement Automatique des Langues*, **49**(1), 41–60.

SAGOT B. & WALTHER G. (2013). Implementing a formal model of inflectional morphology. In C. MAHLOW & M. PIOTROWSKI, Eds., *Third International Workshop on Systems and Frameworks for Computational Morphology*, volume 380 of *Communications in Computer and Information Science*, p. 115–134, Berlin, Allemagne : Humboldt-Universität Springer.

STEIN A. (2014). Parsing heterogeneous corpora with a rich dependency grammar. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Islande.

A. STEIN & AL., Eds. (2008). *Nouveau Corpus d'Amsterdam. Corpus informatique de textes littéraires d'ancien français (ca 1150-1350), établi par Anthonij Dees (Amsterdam 1987), remanié par Achim Stein, Pierre Kunstmann et Martin-D. Gleßgen*. Stuttgart, Allemagne : Institut für Linguistik/Romanistik.

STEIN A. & PRÉVOST S. (2013). Syntactic annotation of medieval texts : the Syntactic Reference Corpus of Medieval French (SRCMF). In P. BENNETT, M. DURRELL, S. SCHEIBLE & R. WHITT, Eds., *New Methods in Historical Corpus Linguistics*, Corpus Linguistics and International Perspectives on Language, p. 275–282. Narr Verlag.

VILLEMONTE DE LA CLERGERIE E. (2005). DyALog : a tabular logic programming based environment for NLP. In *Proceedings of 2nd International Workshop on Constraint Solving and Language Processing (CSLP'05)*, Barcelone, Espagne.

VILLEMONTE DE LA CLERGERIE É., SAGOT B. & SEDDAH D. (2017). The ParisNLP entry at the ConLL UD Shared Task 2017 : A Tale of a #ParsingTragedy. In *Conference on Computational Natural Language Learning*, Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies, p. 243–252, Vancouver, Canada.

# Étude de l'apprentissage par transfert de systèmes de traduction automatique neuronaux

Adrien BARDET[1] Fethi BOUGARES[1] Loïc BARRAULT[1]
(1) LIUM, adresse, 72000 Le Mans, France
prénom.nom@univ-lemans.fr

RÉSUMÉ

L'apprentissage par transfert est une solution au problème de l'apprentissage de systèmes de traduction automatique neuronaux pour des paires de langues peu dotées. Dans cet article, nous proposons une analyse de cette méthode. Nous souhaitons évaluer l'impact de la quantité de données et celui de la proximité des langues impliquées pour obtenir le meilleur transfert possible. Nous prenons en compte ces deux paramètres non seulement pour une tâche de traduction "classique" mais également lorsque les corpus de données font défaut. Enfin, il s'agit de proposer une approche où volume de données et proximité des langues sont combinées afin de ne plus avoir à trancher entre ces deux éléments.

ABSTRACT

**Study on transfer learning in neural machine translation**

Transfer learning is a solution to learn neural machine translation systems when dealing with low ressourced languages pairs. In this paper, We propose an analysis of transfer learning. We want to assess the correlation between data quantity and languages proximity to improve the transfer. We compare these parameters for transfer learning in a classical context and a low resource context. Finally, we want to propose an approach where data quantity and languages proximity are combined so that we do not have to choose between these two elements.

MOTS-CLÉS : apprentissage par transfert, traduction automatique neuronale, quantité de données, proximité des langues.

KEYWORDS: tranfer learning, neural machine translation, data quantity, languages proximity.

## 1 Introduction

Les performances des systèmes de traduction automatique neuronaux évoluent rapidement, cependant cela ne se verifie pas lorsque peu de données sont disponibles. L'apprentissage par transfert est une voie intéressante pour pallier à ce problème. Il consiste à entraîner un modèle sur une tâche bien dotée (modèle "parent"), puis le réutiliser pour l'apprentissage d'un modèle "enfant" (en remplacement d'une initialisation aléatoire). L'objectif est de capitaliser sur les représentations apprises par le système "parent". Ce transfert améliore généralement les résultats du système enfant par un transfert de connaissances apprises par le système parent (Zoph *et al.*, 2016).

Dans cet article, nous analysons différents critères ayant un impact sur l'apprentissage d'un système de traduction automatique neuronal par transfert. Dans notre cas, un premier système est entraîné sur une paire de langues puis réutilisé pour entraîner un second système pour la paire de langues ciblée.

Nous nous intéressons aux différents paramètres qui entrent en compte lorsque l'on emploie ce genre de technique, notamment les caractéristiques des données utilisées pour apprendre le système qui sert de base à notre transfert. En effet, plusieurs travaux portent sur les quantités de données utilisées ainsi que la proximité des langues mises en jeu, et les conclusions divergent (Kocmi & Bojar, 2018; Dabre *et al.*, 2017).

Nous cherchons donc à analyser les configurations de données pour l'apprentissage par transfert afin de déterminer les paramètres pertinents pour obtenir les meilleures performances.

## 2 Travaux Connexes

Certaines grandes avancées techniques ont permis au systèmes neuronaux de devenir l'approche la plus efficace pour la traduction automatique (Sutskever *et al.*, 2014; Bahdanau *et al.*, 2014).

Actuellement, les systèmes neuronaux nécessitent une grande taille de corpus d'entraînement afin d'obtenir de bonnes performances, ce qui, par définition, pose problème pour les paires de langue peu dotées. Les systèmes de traduction automatique statistique à base de segments sont alors une alternative pertinente (Koehn & Knowles, 2017).

Plusieurs approches de traduction automatique multilingue ont été utilisées pour traduire des textes dans des paires de langues peu dotées (Lakew *et al.*, 2018; Gu *et al.*, 2018; Johnson *et al.*, 2016). L'utilisation d'encodeurs et de décodeurs universels ont permis à Johnson *et al.* (2016) de concevoir un système apprenant en parallèle de nombreuses paires de langues et obtenant de meilleurs résultats, notamment pour les langues moins dotées. Des symboles spécifiques (ex : <2es>) sont alors utilisés afin de contrôler la langue en sortie du décodeur universel. Ce genre de modèle permet même de traduire dans des paires de langues non vues pendant l'entraînement (on parle alors de *zero-shot learning*). Cependant, les performances dans de tels cas restent relativement faibles. L'apprentissage par transfert peut combler ce manque en se basant sur un système parent (appris sur une grande quantité de données) qui sert de base pour apprendre un système enfant (Zoph *et al.*, 2016). Le transfert s'opère plus efficacement lorsque des langues sont partagées entre le parent et l'enfant. Dans cette direction, Dabre *et al.* (2017) met en avant l'importance de la proximité des langues pour que le transfert soit de meilleure qualité. Ces observations sont contredites dans Kocmi & Bojar (2018) où de meilleurs résultats sont obtenus avec des paires de langues plus éloignées mais mieux dotées.

Le choix du niveau de représentation des mots est une donnée importante. Nguyen & Chiang (2017) ont montré que l'utilisation de symboles sous-lexicaux partagés entre les langues des modèles parent et enfant, permettent d'augmenter les performances du transfert. Nous utilisons cette méthode en explorant différentes quantités de symboles (i.e. différentes tailles de vocabulaire).

Les travaux présentés dans cet article étendent ceux de Kocmi & Bojar (2018) sur plusieurs points. Tel Kocmi & Bojar (2018), nous cherchons à évaluer les performances du système enfant en fonction des données utilisées dans le système parent, selon les critères de proximité de langue et de quantité de données. Dans cette étude, nous considérons également un système parent constitué d'un encodeur universel (entraîné sur plusieurs langues). Nous nous interrogeons sur les différents choix à effectuer pour les pré-traitements des données et les paramètres du modèle de traduction et nous tenterons de déterminer la meilleure configuration.

L'objectif est de mieux comprendre la corrélation entre l'impact de la quantité de données et celui de

la proximité des langues sur les performances du système enfant. Nous verrons que nos expériences contredisent certaines conclusions des articles précédemment cités.

# 3 Données

Notre objectif est d'avoir les meilleurs résultats possibles pour la paire de langue estonien-anglais. Nous disposons de 2.5 millions de phrases parallèles pour cette paire de langue. Bien qu'on ne puisse pas considérer cela comme une paire de langue sous dotée, cette quantité reste faible pour apprendre un système obtenant de bons résultats.

Nous allons utiliser l'apprentissage par transfert, nécessitant des paires de langues additionnelles. Nous utilisons les données présentées dans la campagne d'évaluation de traduction automatique WMT2018 (Bojar *et al.*, 2018).

Pour évaluer l'impact de la proximité des langues dans le système parent, nous utilisons deux paires de langues différentes. L'une est une paire de langues proche ; nous avons choisi la paire finnois vers anglais car cette langue est proche de l'estonien, ce sont toutes deux des langues finno-ougriennes. Nous disposons de 5 millions de phrases parallèles en finnois-anglais. L'autre paire de langues que nous avons choisie est l'allemand vers l'anglais : l'allemand est une langue germanique plus éloignée du finnois et de l'estonien. En revanche, pour la paire allemand-anglais nous disposons de 40 millions de phrases parallèles, ce qui constitue un corpus de choix. Nous voulons découvrir si cette différence significative de quantités permettra à un système parent allemand-anglais de fournir un transfert au système enfant aussi efficace qu'avec un système parent finnois-anglais.

## 3.1 Pré-traitement de données

Afin de préparer nos données nous passons par plusieurs phases de pré-traitement des corpus. Nous utilisons des unités sous-mots SPM (Kudo & Richardson, 2018). Les systèmes utilisant des unités sous-mots forment l'état de l'art actuel en traduction automatique neuronale. Cela nous permet aussi un transfert plus important entre le parent et l'enfant corrélé au nombre de sous-mots en commun (Nguyen & Chiang, 2017).

Deux modèles d'unités SPM séparés sont appris. Le premier sur les langues sources mises en jeu dans les systèmes parent et enfant, et le second sur la langue cible (anglais). Les deux vocabulaires source et cible correspondant sont créés à partir des données tokenisées à l'aide des modèles précédents.

En entrée de nos systèmes, les langues changent lorsque nous passons de l'apprentissage du système parent à celui de l'enfant. Nous prenons cela en compte en entraînant des modèles de sous-mots pour la partie source avec les données utilisées pour apprendre le système parent et enfant. Le but est de ne pas avoir à modifier le vocabulaire lors de la transition parent/enfant. Afin de ne pas bruiter notre système, nous retirons les phrases de moins de 3 sous-mots et de plus de 100 sous-mots.

Et enfin, nous ne conservons dans nos vocabulaires que les unités sous-mots apparaissant au moins 5 fois dans nos corpus d'entraînement et faisons correspondre les autres à une unité inconnue : <unk>. Ce traitement est nécessaire dans notre cas, puisque SPM ne peut garantir la couverture exhaustive d'un corpus. Nous n'utilisons pas de tags comme dans Johnson *et al.* (2016) pour favoriser le transfert entre les langues proches car, du fait de leur proximité, ces dernières, partagent nécessairement des

sous-mots.

# 4 Architecture

Pour réaliser nos expériences nous nous somme basés sur le principe d'apprentissage par transfert trivial de Kocmi & Bojar (2018). Le principe est d'utiliser une architecture qui ne change pas entre l'apprentissage du système parent et du système enfant. Seules les données d'apprentissage sont changées pour passer de l'apprentissage du parent à l'enfant. Nous utilisons une architecture bout en bout de type encodeur/décodeur standard avec mécanisme d'attention en traduction automatique (Sutskever *et al.*, 2014; Bahdanau *et al.*, 2014). Cette architecture est composée d'un encodeur bi-directionnel et d'un décodeur à base d'unités récurrentes à portes, Bi-GRU (Cho *et al.*, 2014) de taille 800. Les plongements lexicaux (*embeddings*) sont de taille 400. Nous appliquons un *dropout* (Srivastava *et al.*, 2014) de 0.3 sur les *embeddings*, sur le contexte avant qu'il soit fourni au mécanisme d'attention et sur la sortie avant le *softmax*. Nous utilisons Adam (Kingma & Ba, 2014) pour optimiser les poids. Les poids sont initialisés d'après He *et al.* (2015). Le taux d'apprentissage est initialisé à 1.10e-4 et la taille d'un batch est de 32. Cette architecture est la seule configuration utilisée pour tous les systèmes présentés dans cet article. Ils ont été implémentés avec le toolkit nmtpytorch [1] (Caglayan *et al.*, 2017).

# 5 Expérimentations

Tous les résultats des systèmes présentés ici sont calculés sur les corpus de développement de la tâche de traduction de *news* de la campagne d'évaluation WMT2018.

Les résultats des systèmes de base ET-EN présentés dans la table 1 serviront de comparaison aux résultats provenant d'un apprentissage par transfert des systèmes enfants. Les écarts de résultats sont faibles et peu significatifs. À noter que le nombre d'unités SPM employé côté source et côté cible est identique.

| Quantité d'unités SPM | ET-EN 2.5M | ET-EN 200k |
|---|---|---|
| 8k | 14.12 | 10.69 |
| 16k | 14.17 | 10.70 |
| 32k | 13.60 | 10.10 |

TABLE 1 – Résultats en %BLEU pour la paire de langue ET-EN sans apprentissage par transfert avec des vocabulaires comprenant seulement des sous-mots provenant des corpus d'entraînement ET coté source et EN coté cible.

Pour l'apprentissage par transfert, nous conservons le vocabulaire entre les systèmes parents et enfants. De ce fait, nous conservons aussi les modèles SPM. La quantité d'unités sous-mots que nous choisissons pour ces modèles a un impact sur la qualité des systèmes appris. Nous pouvons voir dans la table 2, que les modèles fondés sur les unités SPM comprenant de l'allemand obtiennent de moins

1. https ://github.com/lium-lst/nmtpytorch

bons résultats que ceux comprenant du finnois pour l'apprentissage d'un système estonien-anglais. Le finnois et l'estonien étant des langues proches, il est vraisemblable qu'ils partagent plus de mots qu'avec l'allemand, ce qui explique les résultats obtenus. De ce fait, ils cohabitent mieux dans le vocabulaire. Cela est confirmé par les résultats du modèle SPM estonien-allemand qui augmentent lorsque le nombre de sous-mots augmente. Alors que pour le SPM estonien-finnois les résultats baissent lorsqu'on utilise 32k unités comparées aux 16k unités précédentes. Il semble donc qu'un plus grand nombre de sous-mots soit plus propice pour le système allemand-estonien alors que 16k unités suffisent pour le système finnois-estonien.

| Quantité d'unités SPM | DE+ET 2.5M | DE+ET 200k | FI+ET 2.5M | FI+ET 200k |
|---|---|---|---|---|
| 8k | 10.64 | - | 14.47 | - |
| 16k | 11.55 | 9.27 | 15.08 | 10.66 |
| 32k | 12.52 | - | 13.87 | - |

TABLE 2 – Résultats en %BLEU pour la paire de langue ET-EN sans apprentissage par transfert avec des vocabulaires comprenant des unités SPM relatives aux paires des systèmes parents.

Tout d'abord, les résultats des systèmes de base dans la table 3 nous donnent une idée des performances obtenues par les systèmes parents dans leurs paires de langues respectives. On observe que les performances du système parent DE-EN utilisant seulement 5M de données sélectionnées aléatoirement sont très inférieures à celles du système utilisant toutes les données disponibles. On peut donc s'attendre à une perte de performance lorsque le système parent est entraîné avec une plus faible quantité de données.

Afin non seulement d'avoir des résultats comparables pour nos systèmes enfants, mais aussi pour respecter le principe de l'apprentissage par transfert trivial, nous avons dû nous limiter à une seule configuration d'architecture. Cette dernière est celle décrite précédemment.

Pour définir de la taille de l'architecture nous avons fait un compromis afin d'avoir une taille assez grande pour apprendre correctement le système parent mais raisonnable pour ne pas sur-apprendre lors de l'apprentissage de l'enfant.

Pour la suite des expériences, nous avons choisi d'utiliser 16k sous-mots car c'est avec cette quantité que nous obtenons les meilleures performances en ET-EN.

| Paire de langue | 40M | 5M |
|---|---|---|
| FI-EN | - | 18.03 |
| DE-EN | 20.41 | 11.11 |

TABLE 3 – Résultats des modèles parents de base en %BLEU.

Dans la table 4, nous présentons les résultats des systèmes ET-EN enfants qui ont été appris sur une base des différents systèmes parents présents dans la table 3.

Nos systèmes montrent une amélioration face au 14.17 %BLEU du système de base ET-EN (voir table 1). Les résultats de ces systèmes sont proches mais nous voyons que, de base, les résultats avec le SPM DE+ET sont moins bons. Au final, le meilleur résultat est obtenu avec le transfert du système

FI-EN.

Nous avons utilisé 5M de phrases parallèles extraites aléatoirement des 40M dont nous disposions pour créer un corpus réduit en allemand. Ce corpus nous permet d'entraîner un système parent et de comparer les performances du transfert avec celui du système parent finnois. Les résultats montrent qu'à quantité de données équivalentes, les résultats diffèrent grandement. Nous expliquons cet écart par la proximité des langues utilisées pour entraîner le système parent. Le finnois, plus proche de l'estonien, offre un meilleur transfert que l'allemand qui est plus éloigné. Kocmi & Bojar (2018) montre que la qualité du système parent est importante pour assurer un bon transfert à un enfant. Les performances plus faibles du parent DE-EN utilisant 5M de données sont une explication possible aux faibles résultats du système enfant appris ensuite.

Nous avons aussi essayé de combiner la proximité du finnois à l'estonien et de profiter de la grande quantité de données provenant de la paire allemand-anglais. Pour cela, nous avons réalisé un système avec encodeur et décodeur universels (Ha *et al.*, 2016) avec le corpus finnois et allemand en source de notre système. Le système universel nous permet de modéliser une ou plusieurs langues dans le système sans avoir à faire évoluer l'architecture. Ainsi, nous obtenons des systèmes enfants estonien-anglais vraiment comparables tout en ayant un système multilingue comme parent. De plus, Johnson *et al.* (2016) a montré que l'apprentissage en parallèle de plusieurs paires de langues avec une architecture universelle a un impact positif sur les résultats de traduction, notamment pour les paires de langues dotées d'une quantité de données réduite. Nous voulons vérifier si c'est aussi le cas pour l'apprentissage par transfert. L'idée avec ce système multilingue est d'assembler les deux caractéristiques les plus importantes pour l'apprentissage par transfert, à savoir la proximité des langues impliquées et la quantité de données disponible. Nous utilisons donc un modèle SPM différent des précédents car il comporte cette fois-ci de l'allemand et du finnois provenant du système parent, en plus de l'estonien du système enfant pour le coté source.

L'hypothèse est qu'en combinant ces deux facteurs nous devrions obtenir un parent qui procurera un meilleur transfert à nos systèmes enfants. Les résultats nous montrent que cela n'est pas aussi évident (cf. table 4) ; les performances sont moins bonnes qu'avec l'allemand-anglais ou que le finnois-anglais comme seul parent. Une explication est que le déséquilibre des quantités de données entre les deux langues source du parent est un obstacle à l'apprentissage d'un parent de bonne qualité. Nous envisageons comme travail futur différentes répartitions de données afin de laisser plus de place au finnois dans le système parent.

| Paire de langue | 45M (40M+5M) | 40M | 5M |
|---|---|---|---|
| FI-EN | - | - | 16.55 |
| DE-EN | - | 16.10 | 10.92 |
| FI+DE-EN | 15.71 | - | - |

TABLE 4 – Résultats en %BLEU des modèles enfants ET-EN avec les différents systèmes parents.

Nous constatons une amélioration grâce au transfert pour le système enfant ET-EN (cf. table 4. Toutefois, nous avons aussi voulu appliquer ce transfert dans le cadre de l'apprentissage d'un enfant où peu de ressources sont disponibles. Nous avons donc simulé ce manque de données en ne prenant que 200k phrases du corpus ET-EN pour apprendre de nouveaux enfants avec toujours les mêmes systèmes parents. Les résultats de la table 5 montrent que lorsque peu de données sont disponibles

pour le système enfant, la proximité des langues est plus importante. Le comportement à quantités de données équivalentes ne change pas lors de la comparaison du parent DE-EN à celui FI-EN. Le système parent DE-EN offre un moins bon transfert que le parent finnois. Le système parent FI-EN surpasse clairement les autres dans cette configuration. Il n'y a pas de changement pour notre parent multilingue qui donne toujours un transfert moins performant que les autres.

| Paire de langue | 45M (40M+5M) | 40M | 5M |
|---|---|---|---|
| FI-EN | - | - | 13.03 |
| DE-EN | - | 11.12 | 7.10 |
| FI+DE-EN | 11.05 | - | - |

TABLE 5 – Résultats en %BLEU des modèles enfants ET-EN avec 200k phrases avec les différents systèmes parents.

# 6 Conclusion

Nous avons présenté une analyse de l'apprentissage par transfert pour la traduction automatique neuronale. Cette analyse contient des expériences se concentrant sur deux aspects importants du transfert, à savoir la quantité de données et la proximité des langues. L'objectif est de déterminer lequel de ces deux facteurs est le plus pertinent pour l'apprentissage par transfert. Nous avons montré que les quantités de données et la proximité des langues ont un impact dès la réalisation des unités sous-mots et des vocabulaires. Ces paramètres sont donc à prendre en compte pour le choix des systèmes parents.

Nos résultats vont dans le sens de ceux obtenus par Zoph *et al.* (2016) et Dabre *et al.* (2017) ; la proximité des langues utilisées pour l'apprentissage par transfert est un critère plus important que la quantité de données. À quantités de données équivalentes, les systèmes parents utilisant des paires de langues proches obtiennent de meilleurs résultats. Il ne faut pas, en revanche, négliger la qualité des systèmes parents en question et prendre cela en compte dans les résultats des systèmes enfants. Cette hypothèse est vérifiée lorsque nous avons une quantité "raisonnable" de phrases parallèles pour apprendre un système de traduction et lorsque nous avons peu de données. Notre approche de système universel combinant une grande quantité de données et des données plus proches n'a pas surpassé les approches classiques. Dans le futur, nous aimerions pousser cette approche en essayant différentes combinaisons de quantités de données pour mieux comprendre l'importance de la répartition des langues dans cette approche universelle.

# Remerciements

Ce travail a été effectué dans le cadre du projet CHIST-ERA M2CR, financé par l'Agence Nationale de la Recherche (ANR) sous le contrat numéro ANR-15-CHR2-0006-01.

# Références


BAHDANAU D., CHO K. & BENGIO Y. (2014). Neural machine translation by jointly learning to align and translate. *CoRR*, **abs/1409.0473**.

BOJAR O., FEDERMANN C., FISHEL M., GRAHAM Y., HADDOW B., HUCK M., KOEHN P. & MONZ C. (2018). Findings of the 2018 conference on machine translation (wmt18). In *Proceedings of the Third Conference on Machine Translation, Volume 2 : Shared Task Papers*, p. 272–307, Belgium, Brussels : Association for Computational Linguistics.

CAGLAYAN O., GARCÍA-MARTÍNEZ M., BARDET A., ARANSA W., BOUGARES F. & BARRAULT L. (2017). Nmtpy : A flexible toolkit for advanced neural machine translation systems. *Prague Bull. Math. Linguistics*, **109**, 15–28.

CHO K., VAN MERRIENBOER B., GÜLÇEHRE Ç., BOUGARES F., SCHWENK H. & BENGIO Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, **abs/1406.1078**.

DABRE R., NAKAGAWA T. & KAZAWA H. (2017). An empirical study of language relatedness for transfer learning in neural machine translation. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, p. 282–286 : The National University (Phillippines).

GU J., WANG Y., CHEN Y., LI V. O. K. & CHO K. (2018). Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 3622–3631, Brussels, Belgium : Association for Computational Linguistics.

HA T., NIEHUES J. & WAIBEL A. H. (2016). Toward multilingual neural machine translation with universal encoder and decoder. *CoRR*, **abs/1611.04798**.

HE K., ZHANG X., REN S. & SUN J. (2015). Delving deep into rectifiers : Surpassing human-level performance on imagenet classification. *CoRR*, **abs/1502.01852**.

JOHNSON M., SCHUSTER M., LE Q. V., KRIKUN M., WU Y., CHEN Z., THORAT N., VIÉGAS F. B., WATTENBERG M., CORRADO G., HUGHES M. & DEAN J. (2016). Google's multilingual neural machine translation system : Enabling zero-shot translation. *CoRR*, **abs/1611.04558**.

KINGMA D. P. & BA J. (2014). Adam : A method for stochastic optimization. *CoRR*, **abs/1412.6980**.

KOCMI T. & BOJAR O. (2018). Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation : Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, p. 244–252.

KOEHN P. & KNOWLES R. (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, p. 28–39, Vancouver : Association for Computational Linguistics.

KUDO T. & RICHARDSON J. (2018). Sentencepiece : A simple and language independent subword tokenizer and detokenizer for neural text processing. *CoRR*, **abs/1808.06226**.

LAKEW S. M., EROFEEVA A., NEGRI M., FEDERICO M. & TURCHI M. (2018). Transfer learning in multilingual neural machine translation with dynamic vocabulary.

NGUYEN T. Q. & CHIANG D. (2017). Transfer learning across low-resource, related languages for neural machine translation. *CoRR*, **abs/1708.09803**.

SRIVASTAVA N., HINTON G., KRIZHEVSKY A., SUTSKEVER I. & SALAKHUTDINOV R. (2014). Dropout : A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, **15**, 1929–1958.

SUTSKEVER I., VINYALS O. & LE Q. V. (2014). Sequence to sequence learning with neural networks. *CoRR*, **abs/1409.3215**.

ZOPH B., YURET D., MAY J. & KNIGHT K. (2016). Transfer learning for low-resource neural machine translation. *CoRR*, **abs/1604.02201**.

# Évaluation objective de plongements pour la synthèse de parole guidée par réseaux de neurones

Antoine Perquin[1] Gwénolé Lecorvé[1] Damien Lolive[1] Laurent Amsaleg[2]
(1) IRISA, 6 rue de Kerampont 22300 Lannion, France
(2) IRISA, 263 Avenue Général Leclerc, 35042 Rennes, France
{antoine.perquin, gwenole.lecorve, damien.lolive,
laurent.amsaleg}@irisa.fr

RÉSUMÉ

L'évaluation de plongements issus de réseaux de neurones est un procédé complexe. La qualité des plongements est liée à la tâche spécifique pour laquelle ils ont été entraînés et l'évaluation de cette tâche peut être un procédé long et onéreux s'il y a besoin d'annotateurs humains. Il peut donc être préférable d'estimer leur qualité grâce à des mesures objectives rapides et reproductibles sur des tâches annexes. Cet article propose une méthode générique pour estimer la qualité d'un plongement. Appliquée à la synthèse de parole par sélection d'unités guidée par réseaux de neurones, cette méthode permet de comparer deux systèmes distincts.

ABSTRACT

**Objective evaluation of embeddings for speech synthesis guided by neural networks.**

The evaluation of embeddings extracted from neural networks is complex. The quality of embeddings is relative to the task it was trained for and the evaluation of this task may be a lengthy and costly process if human annotators are involved. Thus, it may be useful to estimate their quality using fast and reproducible objective measures on auxiliary tasks. This paper introduces a generic method to estimate the quality of an embedding. This method is applied to speech synthesis based on unit selection guided by neural networks and allows to compare two systems.

MOTS-CLÉS : Plongements, Évaluation objective, Synthèse de parole.

KEYWORDS: Embeddings, Objective evaluation, Speech synthesis.

## 1 Introduction

La capacité de plongement des réseaux de neurones est souvent utilisée pour obtenir une représentation alternative de données. Extraire la sortie d'une couche cachée d'un réseau de neurones permet d'obtenir une représentation du vecteur d'entrée influencée par la tâche d'entraînement. La qualité d'un plongement est relative à une tâche donnée et l'évaluation de cette tâche peut être un procédé long et onéreux s'il y a besoin d'annotateurs humains.

En particulier, en synthèse de parole, l'évaluation d'un système est effectuée à l'aide de tests d'écoutes subjectifs. Afin d'obtenir des résultats rapidement et d'augmenter la reproductibilité des expériences, différentes mesures objectives existent. Cependant, ces mesures ne sont en général pas corrélées avec les résultats des tests d'écoutes et ne mesurent pas directement la qualité des plongements.

Cet article est un travail exploratoire visant à déterminer la qualité d'un plongement dans le cadre de la synthèse de parole. Nous proposons une méthode générique consistant à comparer un plongement dont on cherche à évaluer la qualité avec ceux issus d'entraînements volontairement défavorables en guise de références basses. La comparaison visuelle de ces plongements permet d'obtenir des critères distinctifs. La mise au point de mesures objectives correspondant à ces critères permet alors de comparer des plongements quelconques. L'application de cette méthode à la synthèse de parole indique qu'un plongement de qualité possède une structure par groupe de phonèmes et que la répartition de ces groupes est informée acoustiquement.

La suite de l'article suit le déroulement suivant : la section 2 présente diverses utilisations et méthodes d'évaluation de plongements, la section 3 présente une méthode générique d'estimation de leur qualité et les plongements utilisés. La section 4 présente l'observation visuelle de ces plongements et la section 5 l'élaboration puis l'utilisation de mesures objectives.

# 2 Travaux liés

La capacité de plongement des réseaux de neurones est utilisée en traitement automatique des langues pour obtenir des plongements de mots (Mikolov *et al.*, 2013). Il s'agit de représenter un mot avec un vecteur de petite dimension (relativement à la taille du vocabulaire) reflétant son contexte. La qualité des plongements peut être évaluée de manière intrinsèque ou extrinsèque (Schnabel *et al.*, 2015). Pour une évaluation extrinsèque, leur qualité est évaluée relativement à une tâche donnée (ex : reconnaissance d'entité nommée (Pennington *et al.*, 2014)) en les utilisant comme attributs d'entrées d'algorithmes d'apprentissage automatique. La qualité du modèle est alors conditionnée par la qualité des plongements utilisés. Pour une évaluation intrinsèque, les relations sémantiques entre mots issues des plongements peuvent être comparés à celles annotées par des humains (Mikolov *et al.*, 2013).

En synthèse de parole, les plongements sont utilisés par les méthodes par sélection d'unités guidée par réseaux de neurones. La méthode par sélection d'unités classique consiste à concaténer des unités de paroles pré-enregistrées afin d'obtenir un signal correspondant à un texte donné (Hunt & Black, 1996). La séquence d'unités à concaténer est choisie au sein d'une base de données comme la séquence qui minimise la somme de deux coûts. Le premier, coût de sélection, indique à quel point une unité dans la base de données est similaire à celle à synthétiser. Ce coût est habituellement défini par des experts linguistes. Le deuxième, coût de concaténation, indique à quel point deux unités consécutives dans une séquence d'unités se concatènent bien. Afin de diminuer la quantité d'expertise linguistique nécessaire pour élaborer des systèmes de synthèse par sélection d'unités, le coût de sélection peut être remplacé par les prédictions d'un réseau de neurones (Merritt *et al.*, 2016), ou la distance euclidienne dans un espace de plongement défini par la couche cachée d'un réseau de neurones (Wan *et al.*, 2017; Perquin *et al.*, 2018). On parle alors de synthèse de parole par sélection d'unités guidée par réseaux de neurones. L'objectif de cet article est la mise au point de mesures objectives de la qualité des plongements servant à guider cette sélection d'unités.

La qualité des plongements pour la synthèse de parole est habituellement évaluée de manière extrinsèque, par des tests d'écoutes. La qualité d'un système seul peut être évaluée à partir de notes arbitraires (Union, 1996) ou par comparaison avec d'autres systèmes (Union, 2003). Cependant, ces tests demandent de nombreux participants pour contre-balancer l'aspect subjectif et être fiables. Afin d'obtenir des indices de qualité de manière rapide et reproductible, différentes mesures objectives existent. Par exemple, la distorsion mel-cepstrale (MCD) permet de mesurer une différence acoustique

entre un signal produit par le système et un signal de référence (Kubichek, 1993). Cette mesure est assimilable à une erreur de reconstruction par le réseau de neurones. Cependant, ces mesures objectives ne sont pas directement des indicateurs extrinsèques de la qualité d'un plongement car elles sont habituellement utilisées pour mesurer la qualité des prédictions du réseau dont sont issus les plongements. L'une des mesures proposée par cette article consiste à les utiliser de manière réellement extrinsèque, en évaluant les prédictions d'un modèle entraîné sur les plongements.

Ce travail présente une méthode générique pour mettre au point des mesures objectives de la qualité d'un plongement. Appliquée à la synthèse de parole guidée par réseaux de neurones, nous proposons différentes mesures de qualité extrinsèque pour des plongements de phones.

# 3 Protocole expérimental

Ce travail est le fruit d'une réflexion sur la qualité d'un plongement et les mesures associées. Nous proposons d'identifier des critères de qualité en comparant un plongement générique avec un plongement jugé mauvais par construction. Cette section présente la méthode employée et les plongements envisagés.

## 3.1 Méthode

La méthode présentée ici est issue de deux constats. Premièrement, il est plus simple de mettre au point une mesure objective distinguant des plongements de qualités très différentes que des plongements de qualités similaires. La méthode propose donc de se concentrer sur l'obtention de mesures permettant de distinguer un plongement quelconque d'un plongement de mauvaise qualité. On peut ensuite vérifier que ces mesures permettent aussi de distinguer des plongements de qualité quelconque. Sous l'hypothèse que la qualité d'un plongement peut être influencée par l'apprentissage du réseau de neurones correspondant, la méthode propose d'entraîner des plongements dans des conditions défavorables afin d'obtenir les plongements supposés de mauvaise qualité. Deuxièmement, il n'est pas toujours intuitif de trouver un critère de qualité d'un plongement. La méthode propose donc de visualiser des espaces de plongements à l'aide d'une méthode de réduction de la dimensionalité. La comparaison visuelle des espaces peut permettre de déduire des critères distinctifs qui serviront de base à la définition de mesures objectives de la qualité d'un plongement.

La méthode proposée pour mettre au point des mesures de la qualité d'un plongement est la suivante :

1. Entraînement de plongements, éventuellement issus de modèles différents, dont on souhaite estimer la qualité ;
2. Entraînement de plongements dans des conditions sous-optimales, ils sont jugés mauvais par construction ;
3. Comparaison visuelle des plongements pour obtenir une intuition de critères distinctifs ;
4. Mise au point de mesures correspondant à ces critères distinctifs ;
5. Comparaison des plongements issus de l'étape 1 avec ceux issus de l'étape 2 afin de vérifier que les mesures objectives correspondent aux critères distinctifs identifiés ;
6. Comparaison des plongements issus de l'étape 1 entre eux pour vérifier que les mesures sont distinctives dans le cas général.

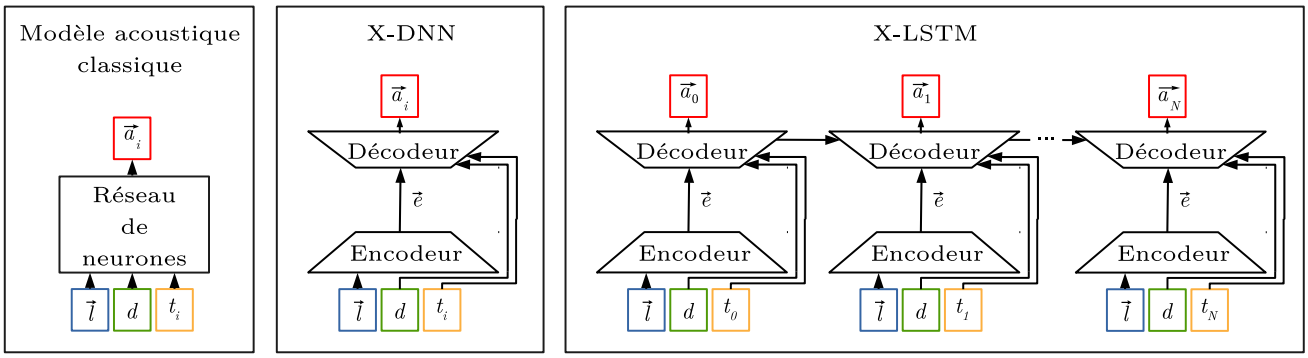



FIGURE 1 – Schématisation d'un modèle acoustique classique et des modèles proposés pour l'extraction de plongements.

Cette étude se concentre sur l'étude de plongements de phones dans le cadre de la synthèse de parole guidée par réseaux de neurones. Les plongements sont obtenus à partir d'un modèle acoustique (schématisé sur la figure 1). Pour un phone de durée $d$ décrit linguistiquement par $\vec{l}$ (identité du phone, de ses voisins, position dans le mot, etc.), pour une trame de position $t_i$ décrite acoustiquement par $\vec{a_i}$ (mel-cepstrum, bande d'apériodicité et fréquence fondamentale), le réseau tente de prédire $\vec{a_i}$ en fonction de $\vec{l}$, $d$ et $t_i$. Afin d'obtenir des plongements de phones plutôt que de trames, $d$ et $t_i$ ne sont fournies qu'après la couche cachée servant de projection.

Sous l'hypothèse que la qualité d'un plongement peut être influencée par l'apprentissage du modèle correspondant, les plongements supposés mauvais seront entraînés dans des conditions volontairement défavorables. Ici, il s'agira d'un sous-apprentissage et d'un sur-apprentissage. Ces mauvais plongements sont choisis car ils peuvent être appris rapidement (peu de données et/ou d'époques d'apprentissage). De plus, en comparaison avec un espace de plongement complètement aléatoire, ils devraient représenter des références basses plus pertinentes. La comparaison visuelle sera effectuée à l'aide d'une projection des plongements par Analyse en Composantes Principales (ACP). Il est important de remarquer que d'autres méthodes de réduction de la dimensionalité pourrait être utilisées. Alors, la visualisation obtenue serait différente et les critères distinctifs pouvant être déduits le seraient aussi.

## 3.2 Modèles

Le premier modèle proposé pour l'extraction de plongements de phones, X-DNN, peut être divisé en deux parties. L'encodeur est composé de 5 couches de dimensions 1024 à 64. Il permet de plonger la description linguistique $\vec{l}$ d'un phone dans un espace vectoriel de dimension 64. Le décodeur est composé de 4 couches cachées de dimensions 128 à 1024 avec une couche finale de dimension 199. À partir d'un vecteur de plongement $\vec{e}$ de phone, de la durée $d$ du phone et de la position $t_i$ d'une trame au sein de ce phone, le décodeur permet de prédire les attributs acoustiques $\vec{a_i}$ associés à la trame. Puisque la fonction d'activation de chaque couche cachée est une tangente hyperbolique, les plongements prennent des valeurs dans l'intervalle $[-1, 1]$.

Le modèle X-LSTM reprend l'architecture de X-DNN en remplaçant la première couche cachée du décodeur par une couche LSTM afin de modéliser les dépendances temporelles au sein d'un phone. En réalité, la présence de cette couche oblige à placer toutes les trames du phone dans un même *batch*, ce qui entraîne une prédiction lissée des $\vec{a_i}$. Malgré des prédictions acoustiques imprécises, les

FIGURE 2 – Visualisation des plongements issus de ref-DNN

FIGURE 3 – Visualisation des plongements issus de sous-DNN

FIGURE 4 – Visualisation des plongements issus de sur-DNN

plongements résultants de X-LSTM permettent une synthèse par sélection d'unités satisfaisante (cf. Tableau 1). Plus d'informations sur les modèles, les attributs utilisés et les performances associées sont disponibles dans (Perquin *et al.*, 2018).

Le jeu de données d'entraînement contient une dizaine d'heures de parole pour un locuteur français masculin professionnel (jeu de données non publié). Cela correspond à 3 300 énoncés soit environ 390 000 phones. La parole est expressive (narration, dialogues joués) et les phrases sont complexes (longues, registre soutenu). Le jeu de donnés est divisé en un sous-ensemble d'entraînement (90%), de test (5%) et de développement (5%).

Les modèles X-DNN et X-LSTM sont entraînés sur la totalité du jeu d'entraînement disponible afin d'obtenir les modèles de références ref-DNN et ref-LSTM (meilleurs modèles sur 100 époques). Les plongements sous-optimaux sont obtenus en entraînant l'architecture X-DNN dans deux cas de mauvais apprentissage. Le modèle sous-DNN est obtenu en entraînant l'architecture sur 256 trames pendant une seule époque. Il s'agit d'un cas de sous-apprentissage. Le modèle sur-DNN est obtenu en entraînant l'architecture sur 256 trames pendant 50 époques. Il s'agit d'un cas de sur-apprentissage. Le nombre de trames d'entraînement pour les modèles sous-DNN et sur-DNN a été choisi pour correspondre à la taille d'un *batch* pour le modèle ref-DNN. Chaque trame correspond à un phone différent, ces modèles ne sont donc appris que sur 256 phones choisis aléatoirement.

# 4 Intuition visuelle

La méthode proposée consiste à définir des mesures de qualité en comparant un plongement donné à un plongement jugé mauvais par construction. Cependant, avant de mettre au point ces mesures, il est nécessaire d'obtenir une intuition sur les critères qui permettent de distinguer les plongements considérés. Nous proposons d'obtenir cette intuition en observant visuellement les plongements, ici par ACP. Chaque point correspond au plongement d'un phone, la couleur de ces points indique le phonème associé.

La figure 2 représente la visualisation par ACP du plongement des phones par le modèle ref-DNN. Graphiquement, il semble que les points de même couleur sont regroupés. Cela signifierait que le plongement de référence permet de grouper les phones associés au même phonème. De plus, dans cet

espace projeté, il semble que les plongements associés aux phones de /p/ et /t/ d'une part et /e/ et /ɛ/ d'autre part sont proches, tandis que ceux des /p/ et des /e/ sont éloignés. Cela indiquerait que l'espace de plongement de référence tient compte d'une similarité acoustique dans la répartition des groupes de phones.

La figure 3 représente la visualisation par ACP du plongement des phones par le modèle sous-DNN. De manière similaire, il semble qu'un plongement sous-appris regroupe les phones par phonème et que la répartition de ces groupes soient informée par une similarité acoustique. En revanche, il semblerait que la distinction entre groupes de phones soit moins claire dans l'espace de plongement sous-appris que dans celui de référence.

La figure 4 représente la visualisation par ACP du plongement des phones par le modèle sur-DNN. Visuellement, il semble impossible de distinguer une structure particulière. En particulier, aucune répartition en groupe de phones n'est remarquable.

La visualisation effectuée n'est en rien une preuve de la qualité d'un plongement ou un reflet exact de sa structure. Cependant, en comparant visuellement le plongement de référence avec les plongements sous-optimaux, deux critères de qualité émergent : une structure de groupe par phonème, une répartition de ces groupes informée acoustiquement. Ces critères sont au final assez intuitifs, mais une méthode de visualisation différente pourrait peut-être permettre de déduire d'autres critères distinctifs.

# 5 Mesures objectives

Cette section présente la mise au point de mesures correspondant aux critères distinctifs issus de l'observation visuelle puis leur application.

## 5.1 Définition des mesures

Le premier critère de qualité proposé est celui de la structure par groupe de phonèmes. Pour évaluer ce critère, on propose de s'intéresser à la notion de plus proches voisins. Soit $e$ le plongement d'un phone quelconque, $e_i$ pour $i \in [1, 100]$ les 100 plus proches voisins de $e$. Si la structure de l'espace de plongement regroupe les phones par groupe de phonèmes, les plus proches voisins du phone plongé doivent correspondre au même phonème. Ce critère peut être mesuré de deux manières complémentaires :

— Comparaison de la classe majoritaire parmi les $e_i$ avec la classe de $e$, en considérant que la classe d'un phone est le phonème associé. On s'intéresse alors à la précision d'une classification par plus proches voisins.
— Mesure du pourcentage des $e_i$ partageant la même classe que $e$. On s'intéresse alors à la pureté du voisinage de $e$.

Le second critère de qualité suggéré par la visualisation des plongements est la répartition des groupes de phonèmes en fonction d'une similarité acoustique. Afin d'éviter l'utilisation d'expertise pour la définition de cette similarité acoustique entre phonèmes, on propose ici de s'intéresser à la mesure du potentiel de prédiction acoustique d'un plongement. Pour une trame d'un phone, un modèle de régression linéaire est entraîné à prédire les coefficients acoustiques $\vec{a_i}$ de la trame en fonction du plongement $\vec{e}$ du phone, de sa durée $d$ et la position $t_i$ de la trame. Ce modèle est entraîné sur le jeu

| | Linguistique | ref-DNN | sous-DNN | sur-DNN | Aléatoire | ref-LSTM |
|---|---|---|---|---|---|---|
| Classification (précision) | **0.972** | 0.952 | 0.893 | 0.882 | 0 | 0.930 |
| Pureté (pourcentage) | **92.9** a† | 92.2 a† | 84.3 | 81.5 | 4.53 | 89.6 |
| MCD (dB) | 6.02 | **5.84** | 6.66 b† | 6.70 b† | 8.21 | 6.20 |
| Test d'écoute ($/10$) (Perquin *et al.*, 2018) | | **7.0** | | | | 6.4 |

TABLE 1 – Mesures objectives pour chaque plongement. † Au sein d'une ligne, suivies de la même lettre, les différences de mesures ne sont pas significatives (0.05)

d'entraînement grâce à la méthode des moindres carrés. La qualité d'un plongement est alors mesurée de manière extrinsèque en calculant la MCD (Distortion Mel Ceptrale) moyenne entre les valeurs prédites et réelles.

## 5.2 Mesures expérimentales

Les mesures définies dans la section 5.1 sont appliquées aux modèles ref-DNN, sous-DNN et sur-DNN. Pour comparaison, elles sont aussi appliquées sur l'espace des descripteurs linguistiques $\vec{l}$, un espace de plongement aléatoire (vecteurs de dimension 64 aléatoirement uniformes sur $[-1, 1]$) et l'espace de plongement défini par ref-LSTM. Les résultats de ces mesures sont rapportés dans le tableau 1.

Par comparaison avec l'espace linguistique, les plongements issus de ref-DNN obtiennent des mesures de classification et de pureté similaire, mais une meilleure mesure acoustique. Ainsi, un plongement correctement entraîné semble conserver la structure de l'espace d'origine, tout en offrant une meilleure capacité de prédiction pour la tâche d'entraînement. En revanche, pour les plongements issus de sous-DNN et sur-DNN, les mesures liées à la conservation linguistique et à la prédiction acoustique sont toutes plus faibles que pour l'espace linguistique. Alors, un plongement mal entraîné semble perdre une partie de la structure de l'espace linguistique et perd même une partie du potentiel de prédiction vis-à-vis de la tâche d'entraînement.

Pour ref-LSTM, les résultats sont légèrement en dessous de ceux de ref-DNN pour toutes les mesures. Cela semble indiquer que les plongements issus de ref-LSTM sont moins bons que ceux issus de ref-DNN. Ces résultats sont cohérents avec les tests d'écoutes effectués dans (Perquin *et al.*, 2018).

# 6 Conclusion

Une méthode est proposée afin d'identifier des critères de qualité d'un plongement, avant de mettre au point des mesures objectives associées. Cette méthode est applicable à tous les plongements, indépendamment de la tâche d'entraînement. Dans le cas de la synthèse de parole, un plongement doit grouper les phones par phonèmes, et ces groupes doivent être répartis selon une similarité acoustique. Les mesures objectives proposées évaluent la qualité du plongement de manière extrinsèque via une classification par plus proches voisins et un modèle de régression linéaire. Cependant, des méthodes de régression plus complexes pourrait permettre de mieux distinguer les plongements. De plus, des méthodes de visualisation autres que l'ACP permettent de mieux conserver les relations de voisinages, ce qui pourrait permettre la découverte de nouveaux critères distinctifs.

# Références

HUNT A. J. & BLACK A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

KUBICHEK R. (1993). Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of the Pacific Rim Conference on Communications, Computers and Signal Processing (CCSP)*.

MERRITT T., CLARK R. A., WU Z., YAMAGISHI J. & KING S. (2016). Deep neural network-guided unit selection synthesis. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the Annual Conference on Advances in Neural Information Processing Systems (NIPS)*.

PENNINGTON J., SOCHER R. & MANNING C. (2014). Glove : Global vectors for word representation. In *Proceedings of the Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

PERQUIN A., LECORVÉ G., LOLIVE D. & AMSALEG L. (2018). Phone-level embeddings for unit selection speech synthesis. In *Proceedings of the International Conference on Statistical Language and Speech Processing (SLSP)*.

SCHNABEL T., LABUTOV I., MIMNO D. & JOACHIMS T. (2015). Evaluation methods for unsupervised word embeddings. In *Proceedings of the Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

UNION I. T. (1996). Methods for subjective determination of transmission quality. *ITUT Recommendation*.

UNION I. T. (2003). Method for the subjective assessment of intermediate quality level of coding systems. *ITUT Recommendation*.

WAN V., AGIOMYRGIANNAKIS Y., SILEN H. & VIT J. (2017). Google's next-generation real-time unit-selection synthesizer using sequence-to-sequence lstm-based autoencoders. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*.

# Exploration de l'apprentissage par transfert pour l'analyse de textes des réseaux sociaux

Sara Meftah[1] Nasredine Semmar[1] Youssef Tamaazousti[3]
Hassane Essafi[1] Fatiha Sadat[2]
(1) CEA LIST, LASTI, Gif-sur-Yvette, France
(2) UQAM, Montréal, Canada
(3) MIT, CSAIL, Massachusetts, États-Unis
{prénom.nom}@cea.fr, ytamaaz@mit.edu, sadat.fatiha@uqam.ca

## RÉSUMÉ

L'apprentissage par transfert représente la capacité qu'un modèle neuronal entraîné sur une tâche à généraliser suffisamment et correctement pour produire des résultats pertinents sur une autre tâche proche mais différente. Nous présentons dans cet article une approche fondée sur l'apprentissage par transfert pour construire automatiquement des outils d'analyse de textes des réseaux sociaux en exploitant les similarités entre les textes d'une langue bien dotée (forme standard d'une langue) et les textes d'une langue peu dotée (langue utilisée en réseaux sociaux). Nous avons expérimenté notre approche sur plusieurs langues ainsi que sur trois tâches d'annotation linguistique (étiquetage morpho-syntaxique, annotation en parties du discours et reconnaissance d'entités nommées). Les résultats obtenus sont très satisfaisants et montrent l'intérêt de l'apprentissage par transfert pour tirer profit des modèles neuronaux profonds sans la contrainte d'avoir à disposition une quantité de données importante nécessaire pour avoir une performance acceptable.

## ABSTRACT

**Exploring neural transfer learning for social media text analysis**

Transfer learning consists in learning a parent neural network on a source problem with enough data, then transferring a part of its weights to represent data of a target problem with few training examples. We present in this paper an approach based on transfer learning to automatically build tools to analyze social media texts by exploiting similarities between texts of a resource-rich language (standard language) and texts of a low-resource language (social media). We conducted experiments on various languages and three Natural Language Processing tasks : Morpho-Syntactic tagging, Part-Of-Speech tagging and Named entity recognition. The obtained results are very satisfactory and show the interest of transfer learning to take advantage of deep neural models without the constraint of having a large amount of data required to obtain an acceptable performance.

MOTS-CLÉS : Apprentissage par transfert, Contenus des réseaux sociaux, Langues peu dotées, Adaptation au domaine, Étiquetage morpho-syntaxique, Reconnaissance d'entités nommées.

KEYWORDS: Transfer learning, Social media content, Low-resource languages, Domain adaptation, Part-Of-Speech tagging, Named entity recognition.

# 1 Introduction

Les méthodes d'apprentissage automatique qui reposent sur les Réseaux de Neurones (RNs) obtiennent des performances qui s'approchent de plus en plus de la performance humaine dans plusieurs applications du Traitement Automatique de la Langue (TAL) qui bénéficient de la capacité des différentes architectures des RNs à généraliser à partir des régularités apprises à partir d'exemples d'apprentissage. En particulier, la structure ordonnée et les dépendances temporelles des données textuelles nécessitent un traitement spécifique. En effet, le contexte joue un rôle important pour identifier le sens d'un mot ou comprendre une phrase dans un document. Pour ce type de tâches, les RNs récurrents (RNNs pour *Reccurent Neural Networks*) et ses variantes ; dont les deux principales : le modèle *Long Short-Term Memory* (LSTM) et sa version simplifiée *Gated Recurrent Units* (GRUs) sont les plus adaptés grâce à leur capacité à conserver en mémoire les informations pertinentes en analysant les mots (ou les caractères) dans un ordre précis. En outre, les RNs à convolutions (CNNs pour *Convolutional Neural Networks*) ont aussi montré leur efficacité pour l'encodage des caractères. Plusieurs études (Jozefowicz *et al.*, 2016) ont montré que les CNNs représentaient l'architecture idéale pour l'extraction et l'encodage des informations morphologiques (racine, préfixe, suffixe, etc.), en particulier pour les langues avec une morphologie riche (Chiu & Nichols, 2015; Ma & Hovy, 2016) comme l'arabe, les langues slovaques, le hindi, etc.

Toutefois, pour être performants, ces modèles neuronaux ont besoin de corpus annotés de taille importante. Par conséquent, uniquement les langues et les domaines bien dotés peuvent bénéficier directement de l'avancée apportée par les RNs, comme par exemple les formes standard de l'anglais, du français, de l'arabe, entre autres. Cependant, la majorité des langues ne sont pas dotées en données d'apprentissage ou ayant des données annotées de très petite taille (Baumann & Pierrehumbert, 2014). Dès lors, les systèmes les plus performants de l'état de l'art pour les langues peu dotées sont ceux fondés sur des règles, construites manuellement dans la majorité des cas. La principale limitation de ces systèmes réside dans leur incapacité à s'adapter à de nouvelles langues et de nouveaux domaines.

Pour pallier au problème de la dépendance des RNs aux données annotées, de nombreux travaux récents s'intéressent à la construction de modèles statistiques et particulièrement des modèles neuronaux pour les domaines peu dotés en exploitant les données annotées des domaines bien dotés (Duong, 2017; Meftah *et al.*, 2018a). Particulièrement, la succession de couches dans les RNs leur permet d'apprendre les connaissances d'une façon hiérarchique, en d'autres termes, des connaissances générales aux couches inférieures et des connaissances spécifiques au problème d'apprentissage aux couches supérieures. Par exemple en traitement d'images les premières couches ont tendance à apprendre les bordures (Yosinski *et al.*, 2014) et les connaissances morphologiques en TAL (Peters *et al.*, 2018), d'où l'intérêt de l'apprentissage par transfert neuronal (TL pour *Transfer Learning*) qui consiste à transférer les connaissances acquises lors de l'apprentissage des problèmes sources aux problèmes cibles.

Nous étudions plus particulièrement dans cet article l'apport du TL par *fine-tuning* pour l'adaptation au domaine, pour contourner le problème de manque de données d'apprentissage dans le domaine des réseaux sociaux (RS). En effet, les performances des outils TAL utilisant des modèles neuronaux appris sur des corpus volumineux annotés manuellement tels que le corpus *Wall Street Journal (WSJ)* de *Penn TreeBank (PTB)* (Marcus *et al.*, 1993) et évalués sur des données du même domaine se rapprochent des performances humaines (97.96% de précision par Bohnet *et al.* (2018)). En revanche, les performances de ces outils chutent fortement lorsque ceux-ci sont appliqués sur des données hors domaine telles que les textes générés par les internautes sur les RS notamment les textes de nature conversationnelle (Twitter, SMS, etc.). Cela est dû principalement aux erreurs linguistiques,

aux incohérences orthographiques, aux abréviations informelles et au style idiosyncratique. De plus, Twitter pose un problème supplémentaire en imposant une limite de 280 caractères pour chaque Tweet. Les expériences menées sur trois tâches : l'étiquetage des séquences des textes des RS : l'annotation en parties de discours (*PoS tagging*), l'annotation morpho-syntaxique (*MS tagging*) et la reconnaissance des entités nommées (*NE recognition*) montrent l'efficacité du *fine-tuning* pour l'annotation des textes des RS.

Le reste de cet article est organisé comme suit : nous commençons dans la section 2 par formaliser l'approche que nous utilisons pour le transfert de connaissances apprises à partir de la forme standard des langues pour l'amélioration de l'annotation des textes des RS. Ensuite, nous décrivons dans la section 3 l'architecture neuronale, les tâches et les corpus sur lesquels nous avons expérimenté notre approche. Nous discutons, par la suite, les résultats que nous avons obtenus dans la section 4. Et finalement, la section 5 conclut notre étude et présente nos travaux futurs.

## 2 Notre approche pour l'annotation des textes de RS

Considérons que nous avons un problème source $\mathcal{P}_s = (\mathcal{D}_s, \mathcal{T}_s)$ et un problème cible $\mathcal{P}_c = (\mathcal{D}_c, \mathcal{T}_c)$, où $\mathcal{D}_s$ et $\mathcal{D}_c$ représentent les domaines source et cible, respectivement, et $\mathcal{T}_s$ et $\mathcal{T}_c$ représentent les tâches source et cible, respectivement. Nous visons à améliorer l'apprentissage de la fonction de prédiction cible à partir des connaissances apprises de $\mathcal{D}_s$ et $\mathcal{T}_s$. Pour cela, nous transférons les connaissances du problème source $\mathcal{P}_s$ au problème cible $\mathcal{P}_c$, tels que $\mathcal{T}_s = \mathcal{T}_c$, $\mathcal{D}_s \neq \mathcal{D}_c$ et $n_s > n_c$ (le domaine source est plus riche en données annotées que le domaine cible). Pour notre cas, le domaine source $\mathcal{D}_s$ représente les textes de la forme standard de la langue et le domaine cible $\mathcal{D}_c$ représente les textes des RS de la même langue, la tâche source et cible $\mathcal{T} = \mathcal{T}_s = \mathcal{T}_c$ est le *PoS tagging*, le *MS tagging* ou *NE recognition*.

Nous utilisons le schéma du *fine-tuning* standard, comme illustré sur la figure 1, nous avons un RNs source $\mathcal{N}_s$ avec un ensemble de paramètres $\theta_s$ répartis en deux ensembles $\theta_s = (\theta_s^1, \theta_s^2)$, et un RNs cible $\mathcal{N}_c$ avec un ensemble de paramètres $\theta_c$ répartis en deux ensembles : $\theta_c = (\theta_c^1, \theta_c^2)$. Le transfert est effectué en trois étapes : (1) $\mathcal{N}_s$ est entraîné sur le problème source avec les données annotées du domaine source $\mathcal{D}_s$. (2) Les poids du premier ensemble de paramètres du $\mathcal{N}_s$ sont transférés au réseau cible $\mathcal{N}_c$ : $\theta_c^1 = \theta_s^1$ (les poids des paramètres $\theta_c^1$ sont initialisés avec les poids appris par le RNs source et ceux des paramètres $\theta_c^2$ sont initialisés aléatoirement) (3) $\mathcal{N}_c$ est affiné sur le problème cible en lançant l'entraînement sur le petit corpus du domaine cible $\mathcal{D}_c$. Notant que le choix du nombre de paramètres (nombre de couches) à transférer dépend de la similarité entre les problèmes source et cible et la disponibilité des données annotées pour le problème cible (Mou *et al.*, 2016).

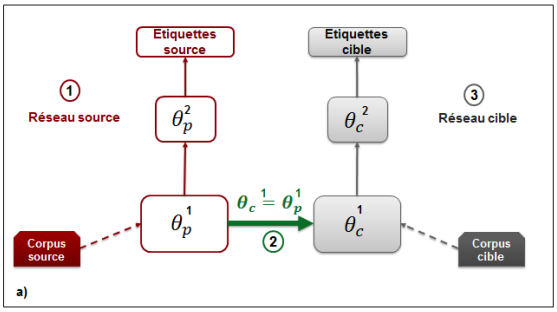



FIGURE 1 – Le schéma de l'adaptation aux domaines par *fine-tuning*.

# 3 Expérimentations

## 3.1 Description des tâches et des jeux de données

Nous expérimentons l'approche discutée sur trois tâches : *MS tagging*, *PoS tagging* et *NE recognition*.

1. Le *PoS tagging* consiste à étiqueter chaque mot dans la phrase avec sa partie de discours (Nom, Adjective, etc.). Pour le domaine source, nous utilisons la partie **WSJ** du PTB (Marcus *et al.*, 1993), annotée avec l'ensemble d'étiquettes du PTB. Et nous évaluons le performances du *fine-tuning* sur trois corpus des Tweets : **TPoS** (Ritter *et al.*, 2011), annoté également avec l'ensemble d'étiquettes du PTB ; **ARK** (Owoputi *et al.*, 2013) annoté avec un ensemble de 25 classes conçues pour les textes des RS ; et **TweeBank**, récemment proposé par Liu *et al.* (2018) et annoté avec l'ensemble d'étiquettes universelles.
2. Le *MS tagging* consiste à étiqueter chaque mot de la phrase avec une étiquette morpho-syntactique, où la partie de discours est enrichie avec les informations morphologiques du mot. Comme illustré sur la figure 2, chaque lettre de l'étiquette représente une catégorie. La première position représente la partie de discours et les autres positions représentent des catégories morphosyntactiques, comme le nombre, le genre, etc. Nous utilisons les données fournies par la compagne d'évaluation *MTT* du **Vardial18** (Zampieri *et al.*, 2018), contenant un corpus de la forme standard et un autre des Tweets pour chacune des langues : Slovène, Croate et Serbe, annotés en *MS tagging*.
3. Pour le *NE recognition*, nous utilisons comme corpus source l'ensemble de données de l'anglais standard de la compagne d'évaluation **CONLL-03** (Sang & De Meulder, 2003) contenant 4 types d'entités nommées : Personne, Organisation, Lieu et *Misc*, la classe des autres entités n'appartenant à aucune des classes précédentes. Concernant le domaine cible, nous utilisons **WNUT-17** de la compagne d'évaluation *Emerging Entity Detection* (Derczynski *et al.*, 2017), contenant 6 types, dont trois communs avec le corpus source : Personne, Organisation et Lieu et trois différents : Produit, Groupe et Travail Créatif.

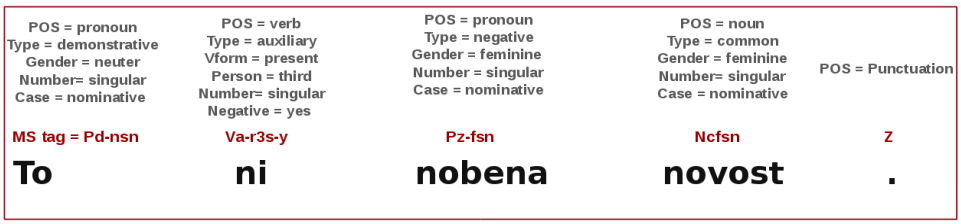



FIGURE 2 – Exemple d'une phrase Slovène annotée en classes morphosyntactiques (To ni nobena novost - "Ce n'est pas une nouveauté.") .

Les statistiques des jeux de données sur lesquels nous avons évalué nos modèles sont résumées dans le tableau 1.

## 3.2 Architecture neuronale

Étant donnée une phrase $S = [w_1, \ldots, w_n]$ de $n$ mots successifs $w_i$, l'objectif d'un étiqueteur est de prédire la classe $c_i \in \mathcal{C}$ pour chaque $w_i$, où $\mathcal{C} \in \mathbb{R}^C$ est l'ensemble de classes. Nous utilisons une architecture neuronale communément utilisée pour l'étiquetage des séquences (Ma & Hovy, 2016; Meftah *et al.*, 2018b). Tout d'abord, chaque mot $w_i$ de la phrase est représenté par $x_i$ : une concaténation de deux plongements de mots hybrides. Le premier est le résultat de l'encodage de tous les caractères du mot avec un LSTMs bi-directionnels (bi-LSTMs) permettant un meilleur traitement

| Tâche | Langue | Domaine | Corpus | # classes | # phrases | # mots |
|---|---|---|---|---|---|---|
| MS | Slovène | Standard | Vardial18 | 1,304 | 27,829 | 586,248 |
| | | RS | | 1102 | 6,670 | 64,108 |
| | Croate | Standard | Vardial18 | 772 | 24,611 | 506,460 |
| | | RS | | 654 | 6,763 | 75,907 |
| | Serbe | Standard | Vardial18 | 557 | 3,891 | 86,765 |
| | | RS | | 589 | 5,884 | 78,616 |
| POS | Anglais | Standard | WSJ | 36 | 67,786 | 1,173,766 |
| | | RS | T-POS | 40 | 787 | 15,000 |
| | | | Ark | 25 | 2,374 | 34,301 |
| | | | TweeBank | 17 | 3,550 | 55,607 |
| NER | Anglais | Standard | CONLL-03 | 4 | 20,744 | 301,418 |
| | | RS | WNUT-17 | 6 | 5,690 | 101,857 |

TABLE 1 – Statistiques des jeux de données utilisés. Pour chaque langue, le corpus de la langue standard est utilisé pour le pré-apprentissage et le corpus du domaine des RS est utilisé pour l'affinement.

des mots hors vocabulaire en capturant leurs caractéristiques morphologiques. Le deuxième est une représentation contextuelle du mot qui permet de capturer sa sémantique. Ensuite, les plongements $[x_1, \ldots, x_n]$ sont fournis dans l'ordre chronologique des mots de la phrase à un extracteur de représentations à base de bi-LSTMs, dont les sorties $[h_1, \ldots, h_n]$ sont acheminées via une couche finale linéaire avec une activation Softmax afin de générer une distribution de probabilité pour chaque classe.

### 3.3 Détails d'implémentation

Les hyper-paramètres utilisés pour notre modèle sont les suivants : nous avons fixé la dimension de la couche de l'embedding des caractères à 50, la dimension de l'encodeur des caractères à base de bi-LSTMs à 100 pour le *NE recognition* et le *PoS tagging* et 200 pour le *MS tagging*. Les plongements de mots sont de dimension 300, ces derniers sont initialisés avec des vecteurs pré-appris et disponibles publiquement. Plus particulièrement, nous avons utilisé les vecteurs Glove (Pennington *et al.*, 2014) pré-appris sur 42 milliards de mots du Web, pour le *PoS tagging* et *le NE recognition*, et les vecteurs pré-appris avec FastText sur des données du web [1] pour le *MS tagging*. Pour la couche bi-LSTMs de l'extracteur de représentations, nous avons fixé sa dimension à 200. Finalement, pour toutes nos expérimentations, l'apprentissage est effectué en utilisant l'algorithme SGD avec momentum et des mini-batches de 8 phrases.

## 4 Résultats expérimentaux et discussion

Nous rapportons dans la première partie du tableau 2 les meilleurs résultats de l'état de l'art pour chaque corpus sur lequel nous avons expérimenté notre approche. L'étiqueteur **ARK** (Owoputi *et al.*, 2013) est basé sur le modèle de Markov d'entropie maximale, utilisant les *brown clusters* et des règles soigneusement conçues à la main. **TPANN** (Gui *et al.*, 2017) est un modèle neuronal, dont l'approche se base sur l'apprentissage antagoniste (Ganin *et al.*, 2016) permettant d'exploiter les Tweets non annotés (1,17M mots) et WSJ. **UH&CU** est le système du *MS tagging* proposé par Silfverberg & Drobac (2018) pour la compagne d'évaluation Vardial 2018. Leur approche est fondée

1. https ://github.com/facebookresearch/fastText/blob/master/docs/crawl-vectors.md

sur un modèle neuronal qui au lieu de considérer une étiquette morphosyntactique comme une entité, génère les différentes catégories morphosyntactiques comme une séquence de caractères en rajoutant un décodeur à base de bi-LSTMs à la sortie du modèle.

Dans la deuxième partie du tableau 2, nous fournissons les résultats expérimentaux de l'utilisation de l'apprentissage par transfert pour l'adaptation des connaissances apprises lors de l'entraînement sur la forme standard d'une langue, au domaine des RS. **Modèle source** consiste à évaluer sur le corpus cible, les performances du modèle appris uniquement sur le corpus source sans adaptation [2]. **Modèle cible** présente les résultats lorsque nous n'entraînons le RNs que sur le corpus des RS (corpus cible), c'est-à-dire, l'initialisation aléatoire de ses paramètres. **Apprentissage par Transfert** montre les résultats de l'approche proposé dans cet article, avec le pré-apprentissage sur le corpus annoté de la langue standard et le *fine-tuning* sur le corpus cible des Tweets.

| **Langue** | Anglais | | | | Sl | Cr | Ser |
|---|---|---|---|---|---|---|---|
| **Tâche** | PoS tagging | | | NER | MS tagging | | |
| **Corpus** | T-Pos | ARK | Tweeb | WNUT | Vardial | | |
| ARK | 90.40 | 93.2 | 94.6 | n/a | n/a | n/a | n/a |
| TPANN | 90.92 | 92.8 | n/a | n/a | n/a | n/a | n/a |
| (Aguilar *et al.*, 2018) | n/a | n/a | n/a | 45.55 | n/a | n/a | n/a |
| UH&CU | n/a | n/a | n/a | n/a | 88.4 | 88.7 | 90 |
| **Modèle source (sans adaptation)** | 76.03 | n/a | n/a | n/a | 70.65 | 78.95 | 73.39 |
| **Modèle cible (sans préapprentissage)** | 87.76 | 90.96 | 91.64 | 41.00 | 82.87 | 84.25 | 85.38 |
| **Apprentissage par Transfert** | **90.70** | **91.81** | **93.00** | **43.57** | **87.83** | **88.30** | **87.06** |

TABLE 2 – Les performances des étiqueteurs de séquence (la précision (%) du *PoS tagging* et du *MS tagging* et la mesure F1 (%) du *NE recognition*) des textes des réseaux sociaux.

En comparant les résultats sans et avec pré-apprentissage, nous constatons que le transfert améliore considérablement les performances pour tous les corpus. En revanche, nous constatons que l'apport est plus important pour *T-PoS* (+3%) et les langues slovaques (+5% pour le slovène). En effet ces corpus partagent les mêmes classes que celles du corpus source utilisé pour le pré-apprentissage, mais cela n'est pas le cas pour *ArK*, *Tweebank* et *WNUT*. Par conséquent, la dernière couche du RNs est pré-apprise pour les expérimentations sur *T-PoS* et les langues slovaques et initialisée aléatoirement pour *ArK*, *Tweebank* et *WNUT*.

Nous traçons dans la figure 3 la courbe des performances de notre modèle sans TL et avec TL sur l'ensemble du développement de Tweebank avec différentes tailles de l'ensemble d'apprentissage des Tweets. Nous pouvons remarqué que le gain apporté par le TL est plus important dans les scénarios où les exemples d'apprentissage du domaine cible sont rares.

Pour mieux comprendre l'impact du pré-apprentissage, la figure 4 présente le pourcentage des prédictions corrigées et celles falsifiées en introduisant le pré-apprentissage du RNs sur le domaine source par rapport à l'initialisation aléatoire. Nous constatons que le pré-apprentissage améliore considérablement les prédictions. Cependant, les falsifications causées par le transfert négatif des régularités spécifiques aux corpus sources (Meftah *et al.*, 2019) réduisent l'apport final du pré-apprentissage. Parmi les erreurs causées par le transfert négatif que nous avons constaté :

1. Les mots avec une première lettre en majuscule ; en effet, dans la forme standard des langues,

2. Notant que cette expérimentation n'est pas possible pour tous les corpus, notamment, les cas où les corpus cibles et sources ne partagent pas le même ensemble de classes (ARK, TweeBank et WNUT)

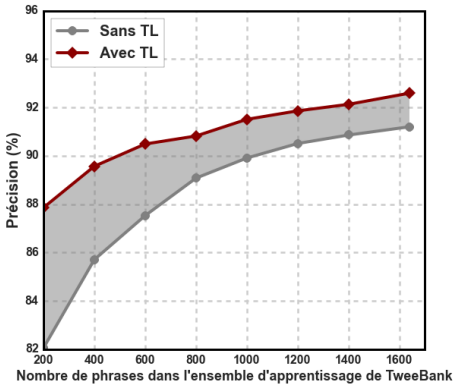



FIGURE 3 – Les performances (sur l'ensemble du développement du TweeBank) avec différentes tailles de l'ensemble d'apprentissage des Tweets.

la première lettre des noms propres est mise en majuscule et le modèle pré-entraîné ne parvient pas à oublier cette régularité qui n'est souvent pas respectée dans les textes des RS.

2. Les mots contenant des ponctuations sont souvent prédits comme *PUNCT*, par exemple : *I'M*.
3. Les verbes *(Did, has, is)* sont prédits comme des auxiliaires dans les cas où ils se trouvent avant un deuxième verbe. Par exemple : *stay informed*.

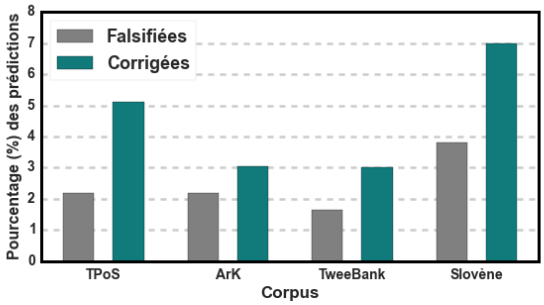



FIGURE 4 – Pourcentage des prédictions corrigées et celles falsifiées en introduisant le pré-apprentissage sur le corpus de la langue standard par rapport à l'initialisation aléatoire.

# 5 Conclusion

Dans cet article, nous avons proposé et implémenté une approche fondée sur l'apprentissage par transfert pour construire automatiquement des outils d'analyse de textes des réseaux sociaux. Cette approche s'appuie sur les similarités entre les textes de la langue standard et les textes des réseaux sociaux relatifs à cette langue. Nous avons montré la validité de notre approche et sa généricité en l'expérimentant sur plusieurs langues et trois tâches d'annotation linguistique. En outre, nous avons constaté que malgré son impact positif, l'apprentissage par transfert est accompagné par un effet de falsification de quelques prédictions des modèles appris sans cet apprentissage.

Nos travaux futurs s'orientent, d'une part, vers l'expérimentation de cette approche sur d'autres langues morphologiquement riches comme l'arabe et ses dialectes, et d'autre part, vers la modélisation des similarités entre la langue standard et ses langues proches en vue d'intégrer cette connaissance externe dans le modèle neuronal pour le forcer à prendre en compte cette connaissance lors de la phase de prédiction.

# Références

AGUILAR G., MONROY A. P. L., GONZÁLEZ F. & SOLORIO T. (2018). Modeling noisiness to recognize named entities using multitask neural networks on social media. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 1401–1412.

BAUMANN P. & PIERREHUMBERT J. B. (2014). Using resource-rich languages to improve morphological analysis of under-resourced languages. In *LREC*, p. 3355–3359.

BOHNET B., MCDONALD R., SIMÕES G., ANDOR D., PITLER E. & MAYNEZ J. (2018). Morpho-syntactic tagging with a meta-bilstm model over context sensitive token encodings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 2642–2652.

CHIU J. P. & NICHOLS E. (2015). Named entity recognition with bidirectional lstm-cnns. *arXiv preprint arXiv :1511.08308*.

DERCZYNSKI L., NICHOLS E., VAN ERP M. & LIMSOPATHAM N. (2017). Results of the wnut2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, p. 140–147.

DUONG L. (2017). *Natural language processing for resource-poor languages*. PhD thesis.

GANIN Y., USTINOVA E., AJAKAN H., GERMAIN P., LAROCHELLE H., LAVIOLETTE F., MARCHAND M. & LEMPITSKY V. (2016). Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, **17**(1), 2096–2030.

GUI T., ZHANG Q., HUANG H., PENG M. & HUANG X. (2017). Part-of-speech tagging for twitter with adversarial neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 2411–2420.

JOZEFOWICZ R., VINYALS O., SCHUSTER M., SHAZEER N. & WU Y. (2016). Exploring the limits of language modeling. *arXiv preprint arXiv :1602.02410*.

LIU Y., ZHU Y., CHE W., QIN B., SCHNEIDER N. & SMITH N. A. (2018). Parsing tweets into universal dependencies. In *NAACL*, volume 1, p. 965–975.

MA X. & HOVY E. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv :1603.01354*.

MARCUS M. P., MARCINKIEWICZ M. A. & SANTORINI B. (1993). Building a large annotated corpus of english : The penn treebank. *Computational linguistics*, **19**(2), 313–330.

MEFTAH S., SEMMAR N. & SADAT F. (2018a). A neural network model for part-of-speech tagging of social media texts. In *LREC*.

MEFTAH S., SEMMAR N., SADAT F. & RAAIJMAKERS S. (2018b). Using neural transfer learning for morpho-syntactic tagging of south-slavic languages tweets. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, p. 235–243.

MEFTAH S., TAMAAZOUSTI Y., SEMMAR N., ESSAFI H. & SADAT F. (2019). Joint learning of pre-trained and random units for domain adaptation in part-of-speech tagging. In *HLT-NAACL*.

MOU L., MENG Z., YAN R., LI G., XU Y., ZHANG L. & JIN Z. (2016). How transferable are neural networks in nlp applications ? *arXiv preprint arXiv :1603.06111*.

OWOPUTI O., O'CONNOR B., DYER C., GIMPEL K., SCHNEIDER N. & SMITH N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. In *NAACL*, p. 380–390.

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

PETERS M. E., NEUMANN M., IYYER M., GARDNER M., CLARK C., LEE K. & ZETTLEMOYER L. (2018). Deep contextualized word representations. *arXiv preprint arXiv :1802.05365*.

RITTER A., CLARK S., ETZIONI O. *et al.* (2011). Named entity recognition in tweets : an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, p. 1524–1534 : Association for Computational Linguistics.

SANG E. F. & DE MEULDER F. (2003). Introduction to the conll-2003 shared task : Language-independent named entity recognition. *arXiv preprint cs/0306050*.

SILFVERBERG M. & DROBAC S. (2018). Sub-label dependencies for neural morphological tagging–the joint submission of university of colorado and university of helsinki for vardial 2018. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, p. 37–45.

YOSINSKI J., CLUNE J., BENGIO Y. & LIPSON H. (2014). How transferable are features in deep neural networks ? In *Advances in neural information processing systems*, p. 3320–3328.

ZAMPIERI M., MALMASI S., NAKOV P., ALI A., SHON S., GLASS J., SCHERRER Y., SAMARDŽIĆ T., LJUBEŠIĆ N., TIEDEMANN J., VAN DER LEE C., GRONDELAERS S., OOSTDIJK N., VAN DEN BOSCH A., KUMAR R., LAHIRI B. & JAIN M. (2018). Language Identification and Morphosyntactic Tagging : The Second VarDial Evaluation Campaign. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, Santa Fe, USA.

# Exploring sentence informativeness

Syrielle Montariol[1, 2] Aina Garí Soler[1] Alexandre Allauzen[1]
(1) LIMSI, CNRS, Univ. Paris-Sud, Univ. Paris-Saclay, F-91405 Orsay, France
(2) Société Générale, 17 Cours Valmy 92043 Puteaux, France
syrielle.montariol@limsi.fr, aina.gari@limsi.fr,
alexandre.allauzen@limsi.fr

RÉSUMÉ

**Explorer l'informativité d'une phrase**

Nous présentons ici une exploration préliminaire du concept d'*informativité* –la quantité d'information qu'une phrase fournit sur l'un des mots qui le compose– et ses usages potentiels pour l'apprentissage de plongements de mots robustes à partir de données en faible quantité. Une mesure d'informativité est prédite à partir d'algorithmes de classification de phrases, que nous comparons à une série de phrases annotées manuellement. Nous concluons que ces deux mesures correspondent à des notions différentes d'informativité. Néanmoins, nos expériences montrent que la prédiction extraite de la classification a un impact sur la qualité des plongements de mots lors de l'apprentissage.

ABSTRACT

This study is a preliminary exploration of the concept of *informativeness* –how much information a sentence gives about a word it contains– and its potential benefits to building quality word representations from scarce data. We propose several sentence-level classifiers to predict informativeness, and we perform a manual annotation on a set of sentences. We conclude that these two measures correspond to different notions of informativeness. However, our experiments show that using the classifiers' predictions to train word embeddings has an impact on embedding quality.

MOTS-CLÉS : Informativité, Plongements de mots, Classification de phrases, Labellisation.

KEYWORDS: Informativeness, Word embeddings, Sentence classification, Data annotation.

## 1 Introduction

Building robust and high-quality word representations is a key step for most NLP tasks. The quality mostly relies on the training corpus (its size and relevance), along with the training criterion and strategy. Some noisy sentences might for instance damage the quality of embeddings, while many sentences do not contribute significantly in improving the representation of a word's meaning (see Table 1 for an example). Therefore, it is advised to estimate embeddings on corpora as large as possible, but this induces many drawbacks : such corpora are difficult to collect for many domains or languages, and the training time increases linearly with the corpus size. Many pre-trained embeddings do exist, built on huge corpora, but they cannot be used in domain-specific tasks.

Consequently, a criterion for sentence selection could be of great help in the context of low-quality or low-volume corpora. The concept at stake is the *informativeness* of a sentence towards a word. This criterion can be defined as follows : *a sentence is informative with respect to one of its words*

*if a person ignoring this word can correctly infer its meaning from this sentence to some extent.* The sentences from the training corpora can be selected depending on their informativeness during embeddings' training. To this end, the informativeness of a sentence needs to be defined and modelled.

The contribution of this paper is threefold. First, we propose a method to automatically get an *artificial* informativeness measure of any sentence towards a target word, using sentence classification. Then, we design a labelling process in terms of informativeness scores to evaluate our models' results. Finally, we give preliminary results on the use of the artificial informativeness measure for the training of word embeddings.

| | |
|---|---|
| A : *I went to the **** | C : *It was good *** in those terms.* |
| B : *I went to the *** to withdraw money* | D : *The aquarium did a blood *** to determine their gender.* |

TABLE 1 – Examples of informativeness disparity between two sentences with respect to a common word. A and B are made-up examples ; C and D are Gigaword (Napoles *et al.*, 2012) sentences with the word *test*.

## 2 Related work

While dealing with low-resource languages and specific domains has received much attention in the NLP community, to the best of our knowledge, this is the first attempt to investigate the potential impact of the notion of informativeness on building meaning representations from limited data. Several attempts at improving or adapting word embeddings to restricted tasks and languages involve making use of morphological information (Luong *et al.*, 2013) ; fine-tuning pre-trained, global-purpose embeddings on a restricted domain (Komiya & Shinnou, 2018; Newman-Griffis & Zirikly, 2018) ; refining them with the help of semantic resources (Faruqui *et al.*, 2015), known as retrofitting ; or using attention mechanisms on contexts to better represent rare words (Schick & Schütze, 2019).

Lai *et al.* (2016) show that, when training word embeddings, using an in-domain corpus specialized in a task is better than having a large, mixed-domain corpus, which can lead to a decrease in performance.

The closest study to the present one is that of Herbelot & Baroni (2017). The authors argue that just as humans do not need many examples to learn the meaning of a word, the word2vec architecture can be adapted to learn a competitive representation of a word from 2-6 occurrences only. This can be critical to build an embedding for rare words or in situations of scarce data. Their experiments involve both Wikipedia definitions –in principle, maximally informative from a human perspective– and naturally-occurring sentences. Given the better performance of the former, they observe that accounting for the informativeness of a sentence can be useful to learn good representations.

A line of work related to informativeness concerns the automatic extraction of sentence examples for lexicographic resources or knowledge bases. Kilgarriff *et al.* (2008) restrict their search with different linguistic criteria designed to help a reader grasp the meaning of a word more easily, such as sentence length and the presence or abscence of rare words, pronouns, or typical collocations. Another approach for finding *knowledge-rich contexts*, as coined by Meyer (2001), consists in finding sentences with knowledge patterns, that is, linguistic expressions that describe the semantic relations of a word, such as "is a kind of" to denote hypernymy (Barrière, 2004).

# 3 Sentence classification for informativeness prediction

According to our definition, if a sentence is very informative with respect to one of its words, a reader who does not know that word can infer its meaning from the sentential context. Intuitively, in such a sentence, a person who knows the word should be able to predict it with high accuracy. We create a task based on this intuition : for a given target word, a model gets as input sentences with a blank at the target position. The model must learn to distinguish sentences that originally contained the target word from those which did not. For instance in Table 1, a model that distinguishes *bank* sentences from *non-bank* sentences should have a higher confidence that the more informative sentence B is a *bank*-sentence. For each sentence, the classifier outputs a probability of belonging to the *bank*-sentences class. Our hypothesis is that this probability is an indicator of the informativeness of the sentence with respect to the word *bank*.

## 3.1 Selection of distractors

In order to make the classification task challenging enough for the models to be effective, we introduce *distractors* : words that share contexts with the target word and could therefore deceive the model. In the first example of Table 1, a good distractor would be a word that, like *bank*, indicates a place (*supermarket*, for example).

For each target word, 10 distractors are selected following Hill & Simha (2016)'s work. First, we collect 3, 4 and 5-grams which include the target word [1] from the 1 million most frequent n-grams in the Corpus of Contemporary American English (COCA). [2] We then search, in the COCA corpus, for words that appear in the extracted n-grams with the same part of speech and position. From this list of potential distractors, words that are synonyms, hyponyms or hypernyms of the target word according to WordNet (Fellbaum, 1998) are removed. Finally, only candidates that have the same or higher frequency than the target word are kept, as calculated from Google unigrams (Brants & Franz, 2006). Ten of the remaining words are randomly chosen to be the target word's distractors.

## 3.2 Classification algorithms

The first classifier relies on a context2vec (c2v) model (Melamud *et al.*, 2016) pre-trained on the ukWaC corpus (Baroni *et al.*, 2009). c2v embeddings are trained with a slot-filling objective and can compute comparable embedding representations of sentential contexts with a blank slot as well as of individual words, where context vectors have a high similarity with those of appropriate fillers. We experiment with logistic regression and a feed-forward neural network with two hidden layers using as input c2v context vectors with a slot at the target word's position.

We compare this to a logistic regression classifier that relies on 3 linguistic-based features to discriminate sentences. Concretely, we use a 3-gram language model [3] and the aforementioned c2v model. For the language model feature, the blank slot is successively replaced with each distractor of a target word. Then, the probability of every resulting sentence is computed. The feature used is the proportion of distractors that, according to the language model, have a higher probability than the target word of

1. With a minimum frequency of 40, 20 and 5 for 3, 4 and 5-grams, respectively.
2. Available at `https://www.ngrams.info/`
3. Available at `http://www.keithv.com/software/giga/` (NVP, 64K words)

filling the slot. The c2v features are the cosine similarity between the target word and the context representation, and the average cosine similarity of every distractor with the context representation.

Several other classification algorithms from the literature were also tested. One with a high accuracy and very low computation time is the FastText classifier (Joulin *et al.*, 2017). It relies on logistic regression ; the sentences are represented with averaging the bag-of-ngrams representation of their words. We put a *"TARGET"* token in place of the target word, a *"NUMBER"* token in place of a number, and –as in the previous models– keep stop words.

# 4 Labelling data

Given a sentence, the classification algorithm outputs a probability of belonging to the class of the target word. According to our hypothesis, this method gives us a measure of informativeness for each pair [*sentence, target word*]. In order to evaluate this measure, two annotators label a set of sentences. Then, we perform an inter-annotator agreement study to select the best methodology and scoring scale.

For a first agreement, the two annotators both label 150 sentences randomly extracted from a portion of the Annotated English Gigaword corpus [4] (Napoles *et al.*, 2012). In each sentence, a target word is randomly masked. It has to be a noun, verb, adjective or adverb, excluding proper nouns and auxiliary verbs. The annotators give two scores to each sentence : $info1$ before seeing the masked word (from 1 to 10, indicating how much they can guess about the target word given the sentence) and $info2$ after seeing this word (from 1 to 10, expressing to what extent they expected seeing the true word). The Spearman correlations between both annotators' scores are relatively low (mean correlation = 0.29 for $info1$ and 0.37 for $info2$), especially for adverbs (correlation = 0.12 for $info2$). Overall, the annotators' remarks show that these measures are very subjective.

Consequently a second labelling agreement is designed, relying on a more explicit measure : $info3$. The range of scores is reduced to be from 1 to 5, and precise scoring guidelines are designed to ensure a common interpretation of scores across annotators. The $info3$ measure answers the question : *How much information does this sentence give about the meaning of the target word ?*

1. The sentence gives no clue about the target word (e.g *I have a ...*)
2. The sentence has at least one element (e.g. *I went to the ....*)
3. The sentence gives some clues about the concept but not very specific (*I went to the ... to speak with the manager.*)
4. The sentence gives a lot of information about the word, but not enough to define it. (*I went to the ... to open a savings account.*)
5. From the sentence, I would be able to write a definition of the word, or this is the only word that could fit here. (*I went to the ... to withdraw money, exchange dollars and ask for a loan.*)

Adverbs are excluded due to their low inter-annotator agreement. Two sentences for each of 50 selected target words are extracted from the same corpus, and manually annotated the same way as for $info2$ : the annotators know which word is the target during labeling.
The global Spearman correlation is 0.331. The annotators rarely disagree on extreme scores, rather on

4. Released by the Linguistic Data Consortium, see `https://catalog.ldc.upenn.edu/LDC2012T21`

medium scores (between 2 and 4). However, the correlation is very low for adjectives (0.080).
To sum up, the second labelling agreement gives a more objective scale, making the annotators usually have close scores, and agreeing on extreme scores. Thus, we keep this process for the rest of the labeling. 20 words are selected out of the 50, excluding adjectives : *call, go, range, carry, charge, coach, hold, return, check, investigator, shot, education, paper, side, figure, post, tell, fire, put, test*. The annotators label together five more sentences for each of the target words, among which we include definitions from WordNet and the online Cambridge Dictionary [5] to ensure the presence of highly informative sentences in the manual annotations. The final evaluation corpus consists of 7 sentences for each of these 20 target words.

# 5 Experiments on test and annotated data

We extract 20,000 sentences for each of the selected target words : 10,000 containing the target word and 1,000 for each of its 10 distractors. 80% are used for training, 10% for development and 10% for testing. We use a different portion of Gigaword than the one used for the manual annotation.
We compare our methods to a simple c2v-based baseline. Given a sentence with a target word, we calculate the similarity between the vectors of all distractors, as well as of the target word, with the c2v context vector of the sentence. The potential fillers are sorted by similarity and the rank of the target word is used as an indicator of informativeness (the more similar, the more informative).

The classifiers described in Section 3.2 are trained on the extracted sentences for each selected target word. Results on the test set are found in Table 2 (first column). The c2v neural model gets the highest accuracy among all classifiers, while the FastText classifier has the lowest. They are then used to make predictions on the manually annotated data. We measure the correlation between the probability of belonging to the target word class and the $info3$ score. Results of this correlation are found in Table 2 (second column). The mean correlation for each classification algorithm is close to zero. Moreover, correlations vary a lot depending on the target word. We conclude that the informativeness measure represented by $info3$ score is not related to the way our classifiers select the most representative sentences for a target word.

Examples of sentences for the target word *tell* can be found in Table 3. The first two are definitions, assigned high informativeness scores by the annotators ; however, the classifier does not recognize the first one as a highly informative sentence. The last sentence is instead classified as a *tell*-sentence with a very high probability even though humans did not find it informative.

| Classifier | Accuracy on test data | Spearman's r on manual data |
|---|---|---|
| Linguistic-based | 0.890 | -0.195 |
| FastText | 0.868 | **0.101** |
| c2v Logistic regression | 0.927 | -0.179 |
| c2v feed-forward NN | **0.946** | -0.162 |
| c2v baseline | 0.759 | -0.070 |

TABLE 2 – Classifiers' results. The first column shows the average accuracy across words on the classification test set. The second one indicates the Spearman correlation of each classifier's prediction with informativeness annotations.

5. Available at `https://dictionary.cambridge.org/`

| Sentence | Human score | Classifier score |
|---|---|---|
| To tell is to let something be known. | 4 | 0.62 |
| To tell means express something in words. | 5 | 0.91 |
| What can I tell him ? | 1 | 0.96 |

TABLE 3 – FastText classifier's probability to belong to the target word's class compared with human annotation for a set of sentences. The target word is the verb "tell".

# 6 Experiments on word embeddings training

We concluded in the previous section that the classifiers output a different kind of informativeness than the human annotations. In this section, we test the effect of the classifiers' informativeness on word embeddings training. We use the probabilities outputted by the Fasttext classifier for this task. Independently of the correlations with the human annotation, this classifier's distribution of outputted probability is the least skewed. The others assign very high probabilities for a large portion of sentences, preventing a clear discrimination between informative and not informative sentences.

Following Herbelot & Baroni (2017), for each target word, we sort the sentences of the test set by probability of belonging to the class of the target word according to the classifier. We select the 250 sentences with lowest and highest probability, and make also a random selection. The target word is replaced by a new token "*target_word_new*" in all sentences. We initialize all weights of a word2vec model (Mikolov *et al.*, 2013) using pre-trained word embeddings. [6] Unknown words are initialized randomly. We fine-tune the pre-trained embeddings on each set of sentences. Then, we compute the similarities between the vector of "target_word_new" and the pre-trained embedding of the target word (gold standard).

Table 4 shows the results of this experiment. $sim\text{-}inf$, $sim\text{-}uninf$ and $sim\text{-}random$ are the similarities computed on the sets of 250 most informative sentences, 250 least informative and 250 random, respectively ; $sim\text{-}inf\&uninf$ includes the 250 most informative sentences and 250 least informative. Looking at the mean differences of each column with $sim\text{-}inf$, we conclude that a low informativeness score of a sentence towards a target word indicates it is less suitable to learn a word embedding ; however, sentences with high probability are not necessarily more helpful than the rest. Moreover, training embeddings on a set of 500 good and bad sentences gives almost the same quality of embeddings as training on a set of only 250 good sentences.

The second part of Table 4 compares the similarities of vectors when trained on 200 random sentences ($sim\text{-}random200$), and when augmenting them with 50 "uninformative" sentences ($sim\text{-}random\text{-}uninf$). The former is usually lower when the $sim\text{-}uninf$ value is low, showing that removing these sentences can improve word representations. However, for words with a high value of $sim\text{-}uninf$, the value $sim\text{-}random\text{-}uninf$ is still higher than $sim\text{-}random200$. Thus, the mean difference is close to zero.

We investigate the reason behind the large disparities among words for this task. We consider two indicators for each word : its frequency in Google Unigrams (Brants & Franz, 2006) and its polysemy in WordNet (Fellbaum, 1998). The Spearman correlation between the word frequency and the difference between $sim\text{-}inf$ and $sim\text{-}random$ is high and negative (-0.6) with a p-value $< 0.05$. Thus, for frequent words, the difference in similarity is low between highly informative sentences

6. Available at `https://code.google.com/archive/p/word2vec/`

and random sentences. On the contrary, in the case of infrequent words, the informative sentences provide better embeddings than the random ones. The correlations with the other columns of Table 4 are not significant. For polysemy, we divide the words into two classes according to the median : the words with less than 11 different senses in WordNet and the words with 11 or more. No significant correlation is observed with the similarity values.

| | **sim-inf** | **sim-uninf** | **sim-inf&uninf** | **sim-rand250** |
|---|---|---|---|---|
| mean | 0.567 | 0.393 | 0.611 | 0.578 |
| diff with 1st col | - | 0.174 | -0.044 | -0.011 |

| **sim-rand200** | **sim-rand&uninf** |
|---|---|
| 0.574 | 0.580 |
| - | 0.006 |

TABLE 4 – The table on the left shows the mean similarity (for the 20 target words) between the pre-trained vector of the target words and the vector trained from selecting only 250 sentences with highest and lowest probability as well as 250 random sentences according to the Fasttext classifier's output. The second table shows the effect that adding 50 low probability sentences to 200 random sentences has on the mean similarity.

# 7 Discussion and future work

In this study, we have introduced the notion of informativeness and proposed an automatic method to predict it for a given set of words. We have performed a manual annotation of informativeness and used it to evaluate our models. Despite their efficiency in classifying sentences, classifiers do not perform well on the manual dataset, suggesting that what the models are learning is different from our definition of informativeness. However, when using the informative and uninformative sentences predicted by the classifier on the task of word embeddings training, we observe that training on uninformative sentences leads in general to lower quality embeddings. Moreover, the average informativeness score of sentences varies a lot depending on the target word : for infrequent target words, informative sentences usually provide a higher gain compared to random sentences.

While not allowing for strong claims about the impact of informativeness on word representations, we believe the results of the present study put forward several interesting questions worth of further research. First of all, it remains to be seen whether the human concept of informativeness can be of help to NLP applications. Several modifications can be introduced to our algorithms, such as the number of distractors or a variety of source corpora ; and annotating a bigger dataset could possibly allow for supervised learning of informativeness.

Another open question is whether embeddings might benefit more from an alternative concept of informativeness. For instance, although definitions are –or should be– informative for humans, they are not very common in most corpora. For this reason, language representations trained on common corpora, like the ones our classifiers rely on, may find them atypical.

With a better understanding of informativeness and automatic predictors, we believe that a study of the features that make a sentence informative for a word would be of great theoretical as well as practical interest, possibly allowing to build target-word-independent predictors. Such features could involve context words sharing a topic with the target, or the degree of polysemy of context words.

# Références

BARONI M., BERNARDINI S., FERRARESI A. & ZANCHETTA E. (2009). The wacky wide web : a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, **43**(3), 209–226.

BARRIÈRE C. (2004). Knowledge-rich contexts discovery. In *Conference of the Canadian Society for Computational Studies of Intelligence*, p. 187–201 : Springer.

BRANTS T. & FRANZ A. (2006). Web 1t 5-gram corpus version 1.1. *Google Inc*.

FARUQUI M., DODGE J., JAUHAR S. K., DYER C., HOVY E. & SMITH N. A. (2015). Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1606–1615 : Association for Computational Linguistics.

C. FELLBAUM, Ed. (1998). *WordNet : An Electronic Lexical Database*. Language, Speech, and Communication. Cambridge, MA : MIT Press.

HERBELOT A. & BARONI M. (2017). High-risk learning : acquiring new word vectors from tiny data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 304–309 : Association for Computational Linguistics.

HILL J. & SIMHA R. (2016). Automatic generation of context-based fill-in-the-blank exercises using co-occurrence likelihoods and google n-grams. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, p. 23–30.

JOULIN A., GRAVE E., BOJANOWSKI P. & MIKOLOV T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 2, Short Papers*, p. 427–431 : Association for Computational Linguistics.

KILGARRIFF A., HUSÁK M., MCADAM K., RUNDELL M. & RYCHLỲ P. (2008). Gdex : Automatically finding good dictionary examples in a corpus. In *Proc. Euralex*.

KOMIYA K. & SHINNOU H. (2018). Investigating effective parameters for fine-tuning of word embeddings using only a small corpus. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, p. 60–67 : Association for Computational Linguistics.

LAI S., LIU K., HE S. & ZHAO J. (2016). How to generate a good word embedding. *IEEE Intelligent Systems*, **31**(6), 5–14.

LUONG T., SOCHER R. & MANNING C. (2013). Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, p. 104–113.

MELAMUD O., GOLDBERGER J. & DAGAN I. (2016). context2vec : Learning generic context embedding with bidirectional lstm. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, p. 51–61 : Association for Computational Linguistics.

MEYER I. (2001). Extracting knowledge-rich contexts for terminography. *Recent advances in computational terminology*, **2**, 279.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, p. 3111–3119.

NAPOLES C., GORMLEY M. & VAN DURME B. (2012). Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, p. 95–100 : Association for Computational Linguistics.

NEWMAN-GRIFFIS D. & ZIRIKLY A. (2018). Embedding transfer for low-resource medical named entity recognition : A case study on patient mobility. In *Proceedings of the BioNLP 2018 workshop*, p. 1–11 : Association for Computational Linguistics.

SCHICK T. & SCHÜTZE H. (2019). Attentive mimicking : Better word embeddings by attending to informative contexts. In *Proceedings of the Seventeenth Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

# Hybridation d'un agent conversationnel avec des plongements lexicaux pour la formation au diagnostic médical

Fréjus A. A. Laleye[1] Antonia Blanié[2] Antoine Brouquet[2] Dan Benhamou[2]
Gaël de Chalendar[1]
(1) CEA, LIST, Laboratoire d'Analyse Sémantique Texte et Image, Gif-sur-Yvette, F-91191
(2) Laboratoire de Formation par la Simulation et l'Image en Médecine et en Santé (LabForSIMS), Faculté de Médecine Paris-Sud, 94275 Le Kremlin Bicêtre
frejus.laleye@cea.fr, gael.de-chalendar@cea.fr, antonia.blanie@aphp.fr, antoine.brouquet@aphp.fr, dan.behnamou@aphp.fr

## RÉSUMÉ

Dans le contexte médical, un patient ou médecin virtuel dialoguant permet de former les apprenants au diagnostic médical via la simulation de manière autonome. Dans ce travail, nous avons exploité les propriétés sémantiques capturées par les représentations distribuées de mots pour la recherche de questions similaires dans le système de dialogues d'un agent conversationnel médical. Deux systèmes de dialogues ont été créés et évalués sur des jeux de données collectées lors des tests avec les apprenants. Le premier système fondé sur la correspondance de règles de dialogue créées à la main présente une performance globale de 92% comme taux de réponses cohérentes sur le cas clinique étudié tandis que le second système qui combine les règles de dialogue et la similarité sémantique réalise une performance de 97% de réponses cohérentes en réduisant de 7% les erreurs de compréhension par rapport au système de correspondance de règles.

## ABSTRACT

**Hybridization of a conversational agent with word embeddings for medical diagnostic training**

In the medical domain, virtual patient or doctor dialogue systems help to train students autonomously to medical diagnosis via simulation. In this work, we exploited the semantic properties captured by distributed word representations to search for similar questions in the dialogue system of a medical conversational agent. We created two dialogue systems that were evaluated on datasets collected during tests with students. The first system based on hand-crafted rules obtains 92% of correct responses on the studied clinical case while the second system that combines rules and semantic similarity achieves a score of 97%. It represents an error reduction of 7% as compared to the rules-only-based system.

MOTS-CLÉS : Agents conversationnels, Chatbots, Médecine, Formation, Similarité sémantique, Plongements Lexicaux.

KEYWORDS: Conversationnal Agents, Chatbots, Medical Education, Semantic Similarity, Word Embeddings.

# 1 Introduction

La pratique du diagnostic médical s'apprend traditionnellement "au lit du malade". Les cours théoriques sont complétés par des stages dans les services hospitaliers. L'étudiant en médecine y observe la pratique des médecins et des internes et pratique lui-même sous le contrôle de ceux-ci. Ce type de formation présente le désavantage de confronter d'emblée l'étudiant en médecine à des situations complexes sans apprentissage pratique (technique et humain) préalable. Il doit gérer en même temps les relations avec des personnes en souffrance et la mobilisation de connaissances complexes et encore incomplètes.

Il apparaît donc utile de pouvoir s'entraîner avant de se confronter à ses premiers patients. Mais pour que cela soit réaliste, il ne peut pas se faire avec des pairs qui joueraient le rôle de patients. Le réalisme serait insuffisant. Il ne peut pas non plus travailler avec des acteurs, le coût serait trop élevé.

Les progrès des outils de réalité virtuelle permettent de plonger à moindre coût l'étudiant dans un environnement réaliste et maîtrisé pédagogiquement. Les chatbots standards permettent de gérer les bases d'un dialogue standardisé entre l'étudiant et le patient virtuel mais sont insuffisants dans le sens ou le moindre écart par rapport au scénario où au lexique prévus les mettent en échec.

Dans cet article, nous montrons comment exploiter des techniques de similarité sémantique fondées sur l'utilisation de plongements lexicaux pour produire un agent conversationnel acceptable pour l'apprentissage des étudiants en médecine.

# 2 Etat de l'art

Les récents progrès des technologies de reconnaissance vocale, du traitement du langage naturel et de l'intelligence artificielle ont conduit à l'émergence des agents conversationnels dans différents domaines de la vie tels que la santé, les finances ou l'éducation. Dans le domaine de la santé, on assiste à l'expansion des patients ou médecins virtuels utilisés pour interagir avec l'homme dans des simulations de scénarios cliniques à des fins de formation, d'éducation ou d'évaluation médicale. Pour la formation en médecine, les apprenants jouent le rôle d'un soignant pour diagnostiquer le patient virtuel et prescrire des traitements avec des scénarios d'entrevue cliniques et pédagogiques validés (Saffari *et al.*, 2014).

Les patients virtuels prennent différentes formes selon les objectifs pédagogiques visés. On les retrouve dans des simulateurs de réalité virtuelle pour la formation en endoscopie (Harpham-Lockyer *et al.*, 2015). Ils sont aussi utilisés, pour enseigner les techniques d'examen oral aux internes en médecine d'urgence (McGrath *et al.*, 2015), pour permettre aux infirmières de développer des compétences en soins aigus telles que l'évaluation et la gestion de la détérioration clinique (Liaw, 2015) et aux stagiaires en médecine de pratiquer les techniques de raisonnement clinique (Close *et al.*, 2015; Kleinert, 2015). Plusieurs programmes en médecine ont déjà intégré les simulations incluant des mannequins pour évaluer la compétence et la confiance des apprenants (Taglieri *et al.*, 2017). Il ressort des différentes évaluations que l'utilisation des patients virtuels fournit des pratiques supplémentaires aux apprenants en dehors des travaux pratiques classiques et améliore leurs performances sur les cas réels. (Smith & Waite, 2017) ont montré que la technologie des patients virtuels peut améliorer les performances des apprenants sur les questions de consultations cliniques.

De nombreuses recherches sont en cours pour améliorer l'efficacité des agents conversationnels

virtuels afin de répondre aux besoins sans cesse croissants des professionnels de la santé. Tous les projets en cours ont pour objectif d'augmenter et d'améliorer l'interaction entre les patients et les médecins (Bioulac *et al.*, 2018; Campillos *et al.*, 2016, 2017). Cette interaction, dans les simulations employant les agents conversationnels dialoguants, se caractérise par : *(i)* le type de technologie (plateforme hébergeant l'agent conversationnel) à savoir les smartphones (Miner *et al.*, 2016; Ireland *et al.*, 2016), les ordinateurs portables ou de bureau (Tanaka *et al.*, 2017; Philip *et al.*, 2017) et les plateformes multimodales (Lucas *et al.*, 2017) ; *(ii)* la stratégie de gestion du dialogue à savoir la stratégie à états finis où le dialogue est une séquence d'étapes prédéfinies (Tanaka *et al.*, 2017; Philip *et al.*, 2017; Lucas *et al.*, 2017) ; la stratégie fondée sur le contenu où le flux de dialogue n'est pas prédéfini mais dépend du contenu fourni par l'utilisateur (Ireland *et al.*, 2016) ; la stratégie fondée sur l'agent où le dialogue s'effectue entre deux agents capables de raisonner (Miner *et al.*, 2016) ; *(iii)* l'initiative dans le dialogue où c'est soit l'utilisateur qui initie la conversation (Miner *et al.*, 2016), soit l'agent virtuel qui conduit la conversation (Tanaka *et al.*, 2017; Philip *et al.*, 2017; Lucas *et al.*, 2017), soit les deux qui peuvent mener la conversation (Fitzpatrick *et al.*, 2017; Ireland *et al.*, 2016).

Dans notre système, les agents conversationnels sont intégrés dans une simulation en Réalité Virtuelle. Ils marient états finis et conversation libre. Certains agents mènent la conversation tandis que d'autres suivent la conversation menée par l'apprenant.

(Jin *et al.*, 2017), réalisant la même tâche de compréhension de questions par un patient virtuel, ont associé un réseau de neurones convolutif et le système ChatScript par un système de combinaison utilisant un classifieur binaire sur des questions collectées lors des séances avec des étudiants en médecine.

# 3 Architecture du système

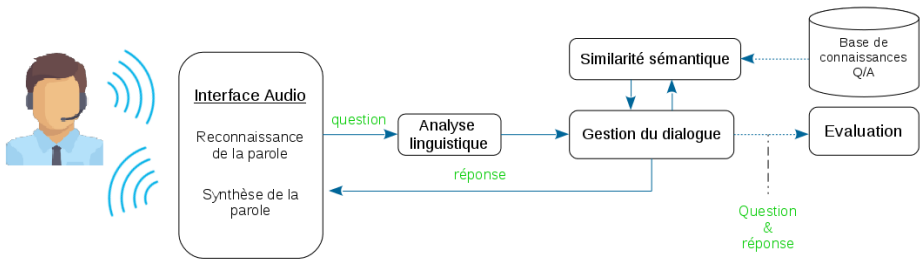



FIGURE 1 – Architecture de l'interaction Apprenant-Agent

La figure 1 présente une vue globale de l'interaction entre l'apprenant et l'agent conversationnel. Le système proposé est conçu pour la consultation d'une urgence chirurgicale abdominale et intègre à la fois la patiente et les médecins (radiologue, chirurgien, etc.) virtuels. La stratégie et l'initiative du dialogue dépendent du profil de l'agent conversationnel. Le système est fondé à la fois sur une stratégie de gestion du dialogue à états finis pour les agents virtuels médecins et une stratégie fondée sur le contenu fourni par l'apprenant pour l'agent virtuel patient. L'architecture présente cinq modules :

— l'interface audio : l'interaction avec l'apprenant se fait via la voix ; l'interface audio constitue le module d'entrée et de sortie du système ; elle intègre un moteur de reconnaissance automatique de la parole (Google Cloud Speech) et un ensemble de générateurs de synthèse vocale (MaryTTS et SVOX Pico) ;

— l'analyse linguistique : l'analyseur linguistique, fondé sur LIMA (Besançon *et al.*, 2010), se charge de filtrer le texte en supprimant les mots vides et de normaliser le texte renvoyé par le moteur de reconnaissance de la parole ;
— la gestion du dialogue : le gestionnaire de dialogue, fondé sur ChatScript (Wilcox & Wilcox, 2013), est un système à base de règles dont les fonctions sont l'interprétation des questions à partir de correspondances de modèles, la génération de réponses et la gestion des modules de recherche de similarité sémantique et d'évaluation. Il se charge, *(i)* ; de retrouver la règle correspondante à la question posée. Les règles sont des sortes d'expression régulières dont les atomes sont les mots. Des classes d'équivalence de mots définissent des concepts. Les règles peuvent prendre en compte un contexte de quelques questions mais le système n'intègre pas de véritable gestionnaire de dialogue. Nous enrichissons ce fonctionnement en permettant au systèmes d'accéder au informations déjà recueillies à propos du patient ou de la patiente. *(ii)* Si la règle est trouvée, ChatScript produit une réponse prévue à l'avance. *(iii)* Sinon la question est envoyée au module de recherche de similarité sémantique qui lui renvoie, à partir de la base de connaissances, une question sémantiquement proche. *(iv)* Il recherche alors à nouveau la règle correspondant à cette question similaire et génère de même la réponse ;
— la recherche de similarité sémantique : l'objectif de ce module est de réduire les limites de compréhension du système à base de règles en assurant une large couverture des questions ; le module de similarité sémantique, fondé sur les plongements lexicaux, calcule une mesure de similarité entre deux questions $q_i$ et $q_j$ (avec $q_i$ la question de l'apprenant et $q_j$ une question déjà associée ou non à des règles de dialogue). Il est composé d'un classifieur neuronal de textes qui, avec comme entrées les plongements lexicaux pré-entraînés, classe une question $q_i$ dans une catégorie $c$ (champ de données cliniques) et d'un calculateur de distances cosinus (voir Eq. 1) entre les vecteurs de plongements lexicaux de la question $q_i$ et l'ensemble des questions de la catégorie $c$. Le résultat est la question de $c$ la plus sémantiquement proche de $q_i$ en terme de distance. La distance est calculée à partir de la somme pondérée des vecteurs mots $(w_{i1}, ..., w_{in})$ de $q_i$ et $(w'_{j1}, ..., w'_{jn})$ de $q_j$ et se présente comme suit :

$$dist(q_i q_j) = 1 - cos\left(\sum_{k=0}^{N} \sigma_k v(w_{ik}); \sum_{k'=0}^{N'} \sigma'_{k'} v(w'_{jk'})\right) \quad (1)$$

où $\sigma_k$ et $\sigma'_{k'}$ représentent respectivement la pondération des mots $k$ et $k'$ avec leurs fréquences inverses en documents.
— l'évaluation : ce module met à la disposition des formateurs des informations issues de la simulation pour une évaluation pédagogique du raisonnement clinique de l'apprenant·e.

# 4 Évaluation et résultats

Dans ce travail, nous présentons les résultats issus de l'évaluation de l'agent conversationnel *patient*. Il s'agit de l'interaction entre l'étudiant et la patiente virtuelle pour une stratégie diagnostique d'une urgence chirurgicale abdominale. L'objectif est d'évaluer la capacité de l'agent conversationnel à fournir des réponses cohérentes aux questions posées par l'étudiant·e.

## 4.1 Description des données

Nous sommes partis des types de questions (*"symptôme : où avez-vous mal ?", "traitement : est-ce que vous êtes sous traitement ?", "antécédent : avez-vous des antécédents médicaux"*) pour écrire les règles ChatScript des différents agents conversationnels. Pour entraîner le classifieur neuronal de textes et le calculateur de distances sémantiques, nous avons collecté les sous-titres de dialogues en français de la série télévisée *ChicagoMed* et les données utilisées par (Campillos *et al.*, 2017) pour la classification automatique de questions médecin-patient. Les questions collectées ont été manuellement annotées suivant les catégories ci-dessus de questions d'une consultation chirurgicale pour des douleurs abdominales. Elles sont utilisées pour construire la base de connaissances du module de similarité sémantique.

Des réponses ont été manuellement associées aux questions afin de construire des jeux de questions/réponses pour l'évaluation globale de l'approche à base de règles et l'approche combinant les règles et la similarité sémantique fondée sur les plongements lexicaux. A la suite des séances de tests effectuées, nous avons recueilli un jeu de données comportant 362 questions. Aussi, nous avons extrait des sous-titres de dialogues de la série *ChicagoMed* un ensemble de 202 questions contenant à la fois des questions d'une consultation pour des douleurs abdominales et des questions qui s'éloignent de ce cas clinique. Le but de ce jeu de données est d'évaluer la capacité de l'agent conversationnel à couvrir un maximum de questions avec le même nombre de règles de dialogue. Chaque jeu de données intègre différentes formulations des mêmes questions. Le tableau 1 présente quelques exemples de question et de leurs reformulations.

| *Type* | *Questions* | *Reformulations* | *Réponses* |
|---|---|---|---|
| Symptôme | Est-ce que vous avez frissonner ? | Avez-vous eu froid ? | Oui j'ai eu froid |
| Sujet de consultation | Qu'est-ce qui vous amène ici ? | Qu'est-ce qui ne va pas madame ? | J'ai très mal au ventre |
| Symptôme | Où se situe la douleur ? | où est-ce que la douleur est localisée ? | J'ai mal au niveau du ventre |
| Personnel | Que faites-vous dans la vie ? | Quel est votre métier | Je suis enseignante |

TABLE 1 – Exemples de reformulations de questions.

## 4.2 Métriques d'évaluation

L'évaluation est réalisée sur la base des types de réponses des deux systèmes. Les métriques pour définir quantitativement les performances de chaque système incluent :

— le nombre de réponses cohérentes équivalant à une bonne compréhension des questions ;
— le nombre de réponses incohérentes équivalant à une mauvaise compréhension des questions ;
— le nombre de réponses *"je n'ai pas compris"* équivalant à une non compréhension des questions.

La cohérence et l'incohérence des réponses ont été jugées par deux médecins experts auteurs du présent article.

## 4.3 Résultats et analyses

La performance de chacun des systèmes est mesurée sur l'ensemble des questions collectées et annotées. Pour les 15 étudiants testeurs, nous avons obtenu un ensemble de 362 questions qui a servi de données pour l'évaluation.

Dans un premier temps, nous avons évalué la capacité des classifieurs (intégrés au module de similarité sémantique) à identifier la catégorie d'une question. Un jeu de 90k questions/réponses a été extrait de l'ensemble des données collectées pour servir de données d'apprentissage et de validation.

Chaque catégorie est traitée comme un problème de classification binaire. Pour l'évaluation, nous avons utilisé une validation croisée en 10 parties. Pour chaque partie, nous avons divisé les données en deux ensembles (apprentissage et validation) avec un ratio de $90/10$. La précision obtenue après apprentissage et évaluation de chaque classifieur est présentée à la figure 2.

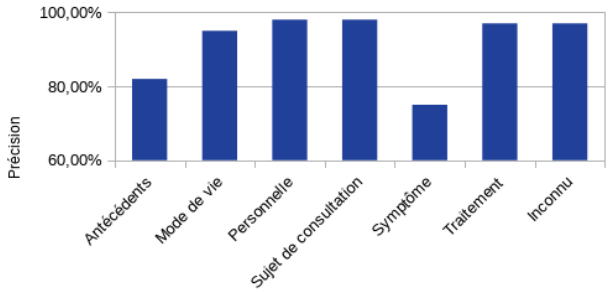



FIGURE 2 – Précision de chaque classifieur de type de questions après la validation croisée.

Nous avons construit le calculateur de distances sémantiques à partir des représentations distribuées des phrases issues des données collectées. Les représentations distribuées des phrases ont été entraînées de façon non supervisée sur un jeu de 45 000 phrases. Le classifieur et le calculateur ont été combinés au système de gestion à base de règles pour une comparaison de performances.

| *Questions* | *Réponses* | *Type de réponses* |
|---|---|---|
| Est-ce que vous avez frissonner | Oui j'ai froid | réponse cohérente |
| Quelle est la couleur de vos urines | Je ne sais pas docteur, j'avais fini de manger | réponse incohérente |
| La pizza était au frigidaire | Je n'ai pas compris. | je n'ai pas compris |

TABLE 2 – Exemple de réponses fournies par le système combiné.

Le tableau 2 présente un exemple de chaque type de réponse. Le tableau 3 montre leurs scores. Pour le corpus issu des tests effectués avec les étudiants, le système combiné réalise une performance de $97\%$ de questions bien comprises (*réponses cohérentes*) en améliorant d'environ $5$ points le score du système à base de règles. En évaluant la capacité des deux systèmes à comprendre une question et à fournir une réponse cohérente ou non, le système combiné a réduit de $7\%$ à presque $0\%$ le taux de réponses *je n'ai pas compris*. Des réponses, parfois incohérentes, sont extraites de la base des questions/réponses du module de similarité lorsqu'il n'y a pas de règles trouvées. Ceci justifie l'augmentation à ($3\%$) du taux de réponses incohérentes. Cette performance vient réduire la frustration de l'étudiant, souvent générée par les réponses *je n'ai pas compris* de l'agent conversationnel lors d'une simulation, qui impacte sur le bon déroulement des interactions dans le sens où l'apprenant doute de la formulation de sa question ou des connaissances de la patiente virtuelle.

| Corpus | Système fondé sur les règles | | | Système combiné (règles + similarité) | | |
|---|---|---|---|---|---|---|
| | *réponses cohérentes* | *réponses incohérentes* | *je n'ai pas compris* | *réponses cohérentes* | *réponses incohérentes* | *je n'ai pas compris* |
| Tests | 92% | 1% | 7% | 97% | 3% | 0% |
| Sous-titres | 65% | 2% | 33% | 94% | 5% | 1% |

TABLE 3 – Évaluation du système fondé sur les règles et de sa combinaison avec la similarité sémantique.

Le taux $33\%$ de questions non comprises par le système à base de règles démontre l'écart entre les questions du corpus issu des sous-titres et le cas clinique étudié. Il montre que les règles de dialogue manuelles éprouvent des difficultés à traiter des questions qui s'éloignent du contexte. L'ajout de la mesure sémantique a réduit de $33\%$ à $1\%$ le taux de réponses *je n'ai pas compris* en augmentant jusqu'à $94\%$ le taux de réponses cohérentes. Ceci montre que, sans ajout de nouvelles règles, la similarité sémantique complète efficacement le système fondé sur les règles de dialogue, rendant le système combiné plus performant.

La combinaison des réseaux de neurones et de ChatScript réalisée par (Jin *et al.*, 2017) a permit d'obtenir une précision de $89,3\%$ en augmentant de plus des deux tiers la précision du réseau de neurones et en réduisant de $47\%$ les erreurs sur ChatScript.

# 5 Conclusion

Nous avons décrit dans cet article un système de dialogue à interface vocale pour la formation des étudiants en médecine au diagnostic des urgences chirurgicales. Notre système marie les capacités de description d'un scénario de dialogue par des règles à la résilience fournie par la similarité sémantique fondée sur des plongements lexicaux. Nous avons atteint un taux de compréhension qui rend le système utilisable par les étudiants. Dans la suite, nous allons concevoir avec les médecins participant au projet les protocoles permettant d'évaluer l'apport pour la formation des étudiants du système global intégrant agent conversationnel et environnement virtuel dynamique.

# Remerciements

Nous remercions Leonardo Campillos Llanos et ses collègues du LIMSI pour nous avoir donné accès à leur corpus d'interactions entre médecins et agent conversationnel. Merci aussi aux étudiants en médecine de l'Université Paris-Sud qui ont bien voulu participer aux tests de notre propre agent conversationnel.

# Références

BESANÇON R., DE CHALENDAR G., FERRET O., GARA F., LAÏB M., MESNARD O. & SEMMAR N. (2010). Lima : A multilingual framework for linguistic analysis and linguistic resources development and evaluation. In *LREC*, p. 3697–3704, Malte.

BIOULAC S., DE SEVIN E., SAGASPE P., CLARET A., PHILIP P., MICOULAUD-FRANCHI J. & BOUVARD M. (2018). Qu'apportent les outils de réalité virtuelle en psychiatrie de l'enfant et l'adolescent ? *L'Encéphale*, **44**(3), 280 – 285.

CAMPILLOS L., BOUAMOR D., BILINSKI E., LIGOZAT A.-L., ZWEIGENBAUM P. & ROSSET S. (2016). Integrating a dialogue system into a virtual patient consultation. In *AMIA Annual Fall Symposium*, Chicago, USA.

CAMPILLOS L., ROSSET S. & ZWEIGENBAUM P. (2017). Automatic classification of doctor-patient questions for a virtual patient record query task. In *16th BioNLP 2017 Workshop*, p. 333–341, Vancouver, Canada.

CLOSE A., GOLDBERG A., HELENOWSKI I., SCHULLER M., DAROSA D. & FRYER J. (2015). Beta test of web-based virtual patient decision-making exercises for residents demonstrates discriminant validity and learning. *Journal of Surgical Education*, **72**(6), 130–136.

FITZPATRICK K. K., DARCY A. & VIERHILE M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot) : A randomized controlled trial. *JMIR Mental Health*, **4**, 19.

HARPHAM-LOCKYER L., LASKARATOS F.-M., BERLINGIERI P. & EPSTEIN O. (2015). Role of virtual reality simulation in endoscopy training. *World Journal of Gastrointestinal Endoscopy*, **7**(18), 1287–1294.

IRELAND D., ATAY C., LIDDLE J., BRADFORD D., LEE H., RUSHIN O., MULLINS T., ANGUS D., WILES J., MCBRIDE S. & VOGEL A. (2016). Hello harlie : Enabling speech monitoring through chat-bot conversations. *Studies in health technology and informatics*, **227**, 55–60.

JIN L., WHITE M., JAFFE E., ZIMMERMAN L. & DANFORTH D. (2017). Combining cnns and pattern matching for question interpretation in a virtual patient dialogue system. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, p. 11–21, Copenhagen, Denmark.

KLEINERT E. A. (2015). Web-based immersive virtual patient simulators : Positive effect on clinical reasoning in medical education. *Journal of Medical Internet Research*, **17**(11), 263.

LIAW E. A. (2015). Designing and evaluating an interactive multimedia web-based simulation for developing nurses' competencies in acute nursing care : randomized controlled trial. *Journal of Medical Internet Research*, **17**(1).

LUCAS G. M., RIZZO A., GRATCH J., SCHERER S., STRATOU G., BOBERG J. & MORENCY L.-P. (2017). Reporting mental health symptoms : Breaking down barriers to care with virtual human interviewers. *Frontiers in Robotics and AI*, **4**.

MCGRATH J., NICHOLAS K., DOUGLAS D., DAVID P. B., SORABH K., DANIEL R. M., ROLLIN N., NICOLE, V. DAVID P. W. & RICHARD N. (2015). Virtual alternative to the oral examination for emergency medicine residents. *The Western Journal of Emergency Medicine*, **16**(2), 336–343.

MINER A., MILSTEIN A., SCHUELLER S., HEGDE R., MANGURIAN C. & LINOS E. (2016). Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA Internal Medicine*, **176**(5), 619–625.

PHILIP P., MICOULAUD-FRANCHI J.-A., SAGASPE P., DE SEVIN E., OLIVE J., BIOULAC & SAUTERAUD A. (2017). Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders. In *Scientific reports*, volume 7, p. 1–7.

SAFFARI Z., TAKMIL F. & ARABZADEH R. (2014). The role of educational technology in medical education. *Journal of Advances in Medical Education & Professionalism*, **2**(4), 183.

SMITH M. & WAITE L. (2017). Utilization of a virtual patient for advanced assessment of student performance in pain management. *Journal of Currents in Pharmacy Teaching and Learning*, **9**(5), 893–897.

TAGLIERI C. A., CROSBY S. J., ZIMMERMAN K., SCHNEIDER T. & PATEL D. K. (2017). Evaluation of the use of a virtual patient on student competence and confidence in performing simulated clinic visits. *American Journal of Pharmaceutical Education*, **81**(5), 87.

TANAKA H., NEGORO H., IWASAKA H. & NAKAMURA S. (2017). Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders. *PLOS ONE*, **12**(8), 1–15.

WILCOX B. & WILCOX S. (2013). Making it real : Loebner-winning chatbot design. In *Arbor*, volume 189, p. a086.

# Inférence des relations sémantiques dans un réseau lexico-sémantique multilingue

Nadia Bebeshina-Clairet Mathieu Lafourcade
Université de Montpellier, LIRMM, France
clairet@lirmm.fr, lafourcade@lirmm.fr

RÉSUMÉ

Les méthodes endogènes se trouvent au coeur de la construction des ressources de connaissance telles que les réseaux lexico-sémantiques. Dans le cadre de l'expérience décrite dans le présent article, nous nous focalisons sur les méthodes d'inférence des relations. Nous considérons, en particulier, les cas d'inférence des relations sémantiques et des raffinements de sens. Les différents mécanismes d'inférence des relations sémantiques y compris dans le contexte de polysémie de termes ont été décrits par Zarrouk (2015) pour le contexte monolingue. À notre connaissance, il n'existe pas de travaux concernant l'inférence des relations sémantiques et des raffinements dans le contexte d'amélioration d'une ressource multilingue.

ABSTRACT

**Inferring semantic relations in a multilingual lexical semantic network.**

Endogeneous methods are important for the graph based knowledge resource building. In the framework of the experiment we describe in the present paper, we focus on inferring new relations. In particular, we consider semantic relation and sense refinement inference. Various mecanisms of semantic relation inference including those with term polysemy have been detailed by Zarrouk (2015) in the monolingual context. To our knowledge, no experimental work have been done in order to elaborate such methods for a multilingual resource enhancement.

MOTS-CLÉS : réseau lexico-sémantique, ressource multilingue, inférence de relations.

KEYWORDS: lexical semantic network, inference, multilinguality.

# Introduction

L'inférence endogène des relations sémantiques représente une moyen intéressant d'enrichissement des ressources lexico-sémantiques multilingues. En effet, la construction de ces-dernières est susceptible de s'appuyer sur des ressources structurées pré-existantes plus facilement disponibles pour les langues dites « riches ». Le contexte endogène offre ainsi la possibilité d'enrichir les partitions consacrées aux langues dites « peu dotées » à partir des relations sémantiques disponibles dans les partitions de langues « riches ». Cet aspect a également été souligné dans (Huang *et al.*, 2002) où l'expérience d'amorçage du WordNet du Chinois grâce à Princenton WordNet (PWN, (Fellbaum, 1998)) est décrite. Après avoir conçu un réseau lexico-sémantique multilingue pour un domaine de spécialité (alimentation), nous nous sommes tournés vers ces méthodes afin d'optimiser la construction de notre ressource.

# 1 État de l'Art, contexte de l'expérience

Pour les bases de connaissance factuelles telles que NELL (Carlson *et al.*, 2010), plusieurs approches d'inférence centrée sur l'équivalence des entités et des relations ont été proposées (par exemple, l'expérience de fusion de plusieurs éditions monolingues de NELL décrite dans (Hernández-González *et al.*, 2017)). Le processus d'inférence endogène a été étudié par Zarrouk (2015) et Ramadier (2016) dans le cadre d'un réseau lexico-sémantique pour le français RezoJDM (Lafourcade, 2007). Leurs méthodes reposent sur l'exploration des relations présentes dans ce réseau afin d'en proposer des nouvelles en suivant des schémas d'inférence par déduction/induction (exploitation des relations taxonomiques), par abduction (exploitation des termes jugés similaires), par raffinement. Gelbukh (2018) introduit un mécanisme d'inférence similaire à celui proposé par ces auteurs pour l'enrichissement d'une base collocationnelle et, notamment, l'inférence par abduction (inférence fondée sur la similarité sémantique acquise à partir de PWN) de nouvelles collocations. L'ancrage socioculturel de l'alimentation détermine les particularités de sa langue (présence de nombreux concepts implicites). Les méthodes d'analyse des textes de cuisine, un des domaines d'application de la ressource construite nécessitent un contexte riche qui se présente notamment comme méta-langage spécifique (approches décrites dans (Tasse & Smith, 2008), (Jermsurawong & Habash, 2015)), structures dynamiques (vecteurs d'état latents), vocabulaires ou ontologies construits à la volée, exploitation des ressources de connaissance sous forme de graphe comme celle construite dans le cadre de nos expériences.

Le réseau lexico-sémantique multilingue avec pivot interlingue ($\text{RLSM}_{\text{PI}}$) constitue le contexte de nos expériences. Le réseau contient 821 781 termes et 2 231 197 relations à l'heure où nous écrivons. Le modèle de $\text{RLSM}_{\text{PI}}$ (figure 1) s'inspire de celui du réseau lexico-sémantique RezoJDM, (Lafourcade 2007 & 2011). Il s'agit d'un graphe *orienté* (où chaque relation possède un terme source et un terme cible), *typé* et *valué* [1]. Ce graphe comporte $k$ sous-graphes correspondant à chacune des $k$ langues couvertes par la ressource (l'anglais, le français, le russe et l'espagnol) et un *pivot interlingue*. Pour éviter les difficultés propres à la construction d'un pivot artificiel dont la nécessité d'aligner $N$ sens simultanément et dans l'impossibilité d'obtenir facilement un modèle d'alignement global sur la base d'un plongement complet du pivot interlingue, le pivot est amorcé comme un pivot naturel en utilisant l'édition anglaise de DBNary (Sérasset, 2012). Il évolue vers un pivot interlingue de façon incrémentale ce qui permet de réduire progressivement le phénomène contrastif artificiel défini par (Sérasset, 2012) comme « une perte d'information discriminatoire liée à une conceptualisation et lexicalisation qui divergent entre les langues » propre à l'utilisation d'un pivot naturel.

# 2 Inférence translingue des relations sémantiques

## 2.1 Principe et déroulement

Dans le cadre du $\text{RLSM}_{\text{PI}}$, il s'agit d'inférer les relations sémantiques dans les partitions qui en contiennent peu grâce aux relations présentes dans les partitions riches de cette ressource. L'exemple simplifié du terme russe пряник pour lequel on souhaite proposer des relations typées *r_has_part* grâce à la partition « fr » du $\text{RLSM}_{\text{PI}}$ permet d'illustrer la mise en oeuvre du processus d'inférence dans le contexte multilingue. Il existe une distinction de sens du terme *pain d'épices* en français qui peut être modélisée au niveau interlingue sous forme de deux raffinement du terme générique `in:`*gingerbread*

1. Ses arcs sont caractérisées par des poids et des annotations

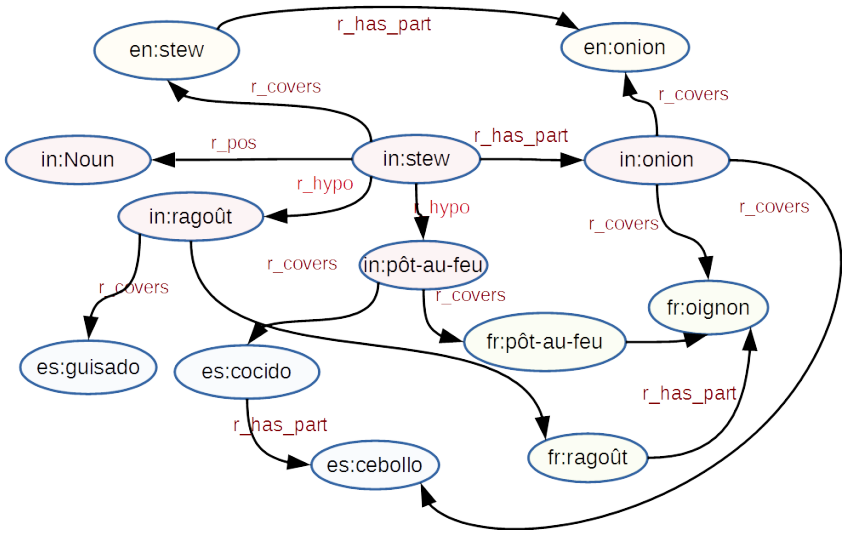



FIGURE 1 – Architecture du réseau lexico-sémantique multilingue avec pivot interlingue. Les éléments des partitions sont reliés uniquement via le pivot interlingue via la relation typée *r_covers*. Le terme interlingue est appelé *terme couvrant* et le terme lexicalisé correspondant est défini comme le *terme couvert*. Les sens d'usage des termes sont appelés *raffinements*.

qui sont `in:`*gingerbread>cake* et `in:`*gingerbread>biscuit*. L'inférence des relations sémantiques se déroule en deux temps. D'abord, elles sont inférées dans le pivot (dans notre exemple, à partir de `fr:`*pain d'épices*vers `in:`*gingerbread* ) à l'aide des termes couvrants les termes-voisins de *pain d'épices* et de ses raffinements tels que `in:`*sugar* $\xrightarrow{r\_covers}$ `fr:`*sucre*, `in:`*ginger* $\xrightarrow{r\_covers}$ `fr:`*gingembre*, etc. Ensuite, les relations du pivot sont inférées dans les partitions à enrichir (dans notre exemple, la partition « ru » : `ru:` пряник $\xrightarrow{r\_has\_part}$ `ru:` сахар, etc.).

Le système d'exploitation du $RLSM_{PI}$ se rapproche du système défini par (Ferber, 1995) comme une architecture fondée sur les *systèmes de production* [2]. L'inférence n'est pas conçue comme un processus séquentiel. Les relations exploitées dans le cadre d'inférence translingue pour identifier les termes dont les relations sémantiques génèrent les prémisses des règles d'inférence sont des relations typées *r_covers*. Elles relient le terme interlingue aux termes qu'il couvre. Par conséquent, aussi bien dans la phase ascendante (langue → pivot) que dans la phase descendante (pivot → langue), nous pouvons supposer qu'il s'agit des termes équivalents. Or, un seul et même terme couvert peut avoir plusieurs termes interlingues couvrants qui peuvent correspondre à plusieurs sens. Le cas inverse est également fréquent. Ainsi, nous considérons la relation typée *r_covers* comme une variante translingue de synonymie potentiellement incomplète. La conséquence de ce positionnement se traduit par le fait que ***en présence de plusieurs termes interlingues couvrants, le cas d'inférence est traité comme s'il s'agissait d'une inférence avec raffinements.*** Ce processus consiste à vérifier la présence des relations sémantiques entre les raffinements du terme et l'extrémité opposée de la relation à inférer. Ainsi, dans le cas du terme `en:`*cake* qui a plusieurs termes couvrants dans le pivot interlingue dont `in:`*cake>sweet food*, `in:`*cake>galette*, `in:`*cake>block* et pour lequel nous souhaitons proposer une relation typée *r_has_part* vers `in:`*ginger* nous pouvons choisir de nous assurer qu'il existe un voisin du terme interlingue (par exemple, un générique tel que `in:`*sweet food*) qui a une lien sémantique (relation ou chemin typé de longueur 2) avec « `in:`*ginger*.

2. *Un système de production est défini par la combinaison d'une base de faits (BF), d'une base de règles de production (BR) et d'un interprète, le moteur d'inférence (MI). Op.cit.* p. 137. Les règles de production ont la forme générale : « si *liste de conditions* alors *liste d'actions* ».

*Dans le cadre monolingue*, le mécanisme d'inférence adapté au cas des termes similaires est l'inférence par abduction. Il s'agit de sélectionner un ensemble de termes similaires à un terme $T$ et de proposer les relations détenues par ces termes à $T$. On considère ainsi les demi-relations partagées par une paire de termes (relations typées de/vers un terme-voisin).

*Dans le cadre multilingue et ascendant* la relation à inférer est considérée comme une instance de règle d'inférence par abduction. On transforme ses termes source et cible en ensembles de termes qui peuvent contenir aussi bien des termes interlingues que lexicalisés. On recherche des « faisceaux » de relations existantes entre ces ensembles de termes. Autrement dit, on récupère tous les termes interlingues et lexicalisés pour les deux termes de la relation considérée. Puis, on explore le voisinage de l'intersection des ensembles de termes ainsi obtenus. Si la cardinalité de l'intersection entre les voisinages typés est suffisante (définie par un seuil [3]), la relation issue de la partition lexicalisée est proposée pour les termes du pivot interlingue. *Dans le cadre multilingue et descendant*, le processus d'inférence s'appuie principalement sur le filtrage logique par triangulation car de multiples termes couverts pour un seul terme couvrant sont possibles. Il s'agit de vérifier la présence d'un chemin typé entre les termes lexicalisés couverts respectivement par le terme interlingue source et par le terme interlingue cible de la relation à inférer. À terme de l'évolution du pivot, l'algorithme d'inférence des relations deviendra celui d'une simple « remontée-descente » et ne nécessitera plus de recours systématique à un mécanisme d'inférence.

## 2.2 Filtrage

Les inférences fondées sur les exemples génèrent un nombre important de relations-candidates qui nécessitent une procédure de filtrage afin de ne pas introduire de bruit (dû à la polysémie des termes) dans la ressource. Nous avons appliqué un *pré-filtrage par parties du discours* car celui-ci, étant une simple vérification de la présence et du poids suffisant [4] de la relation typée *r_pos* permet de réduire le temps de calcul des relations candidates. La plupart des relations sémantiques considérées dans les expériences que nous décrivons relient deux termes dont la partie de discours est « nom ». D'autres relations telles que *r_carac* (caractéristiques typique) mais aussi les relations actantielles [5] dans le cadre des langues flexionnelles telles que le russe imposent des contraintes morpho-syntaxiques qui peuvent être exploitées. Nous avons introduit un *filtrage statistique* car les relations du $\text{RLSM}_{\text{PI}}$ peuvent être analysées en considérant leur *nombre*, leur *poids* et leur *origine*. Le *poids* $w$ correspond à la *force d'association* et s'applique aux relations issues des ressources construites par peuplonomie comme, par exemple, RezoJDM où il est proportionnel à la fréquence avec laquelle les termes source et cible de la relation sont associés par les joueurs [6]. Le cas échéant, il est fixé par défaut. L'*origine* est une liste de chaînes de caractères qui désignent les processus ou les ressource qui ont fourni la relation. Dans le cadre du filtrage, nous avons introduit l'information sur la confiance accordée à l'origine d'une relation (ressources de connaissance, processus endogènes) sous forme d'un ensemble d'indices de confiance $\psi = \{i_1, i_2, ...i_n\}$ où $i_j \in [0; 1]$ ainsi que la cardinalité de l'ensemble des demi-relations partagées par les termes $\phi$. La fonction de filtrage se calcule pour un poids positif ou négatif de la relation $w \in Z$ et $|\psi| > 0$ comme suit :

$$f(r) = \phi \times \frac{w}{Max(\psi) \times \log(|\psi|)}$$

3. Ce seuil a été empiriquement fixé à 3.
4. Le poids suffisant fixé empiriquement est un poids $w \geq 25$.
5. Relations dites « prédicat - arguments » : *r_object* (patient typique), *r_instr* (instrument typique), etc.
6. Dans le contexte des *GWAP* ( *Games With a Purpose*), « jeux avec un but ».

Un score combinant la cardinalité et le maximum de $\psi$ ainsi que le poids est utilisé lorsqu'il ne s'agit pas des termes supposés similaires. Outre le filtrage statistique, le filtrage logique par triangulation peut s'appliquer.

## 2.3 Expérimentation

Les expérimentations ont été conduites sur l'ensemble des relations sémantiques et sur l'ensemble des langues du $RLSM_{PI}$. Le chiffrage des expérimentations s'appuie sur les informations telles que le nombre de relations dans la partition d'origine (**#orig**), le nombre de relations candidates (**#cand**) soit le nombre de relations qui vérifient les prémisses d'une règle d'inférence, la productivité de l'algorithme (**prod**) qui correspond au nombre de candidats par rapport à celui des relations présentes dans la partition d'origine, le nombre de relations acceptées (**#acc**) soit le nombre de relations candidates qui permettent de fournir la conclusion et qui subsistent après la procédure de filtrage, le pourcentage des relations acceptées (**%acc**) et la précision (**pr**) évaluée manuellement sur un échantillon de 500 relations par type de relation. **rang** a été introduit pour exprimer le rapprochement de la productivité « idéale » qui se situerait autour de 100% ». La partition **fr** est la plus riche en termes du nombre de types de relations. L'inférence de certaines relations taxonomiques et méronymiques affiche une productivité assez faible due aux intersections entre les ressources déjà intégrées dans le $RLSM_{PI}$ dont WordNet (Fellbaum, 1998)) ainsi que la nature non encore interlingue du pivot.

| **type** | *lang* | **#orig** | **#cand** | **prod** | **#acc** | **%acc** | **rang** | **pr** |
|---|---|---|---|---|---|---|---|---|
| *r_isa* | fr | 148 409 | 24 379 | 16 % | 13 886 | **57 %** | 1 | 72 % |
| | en | 314 452 | 50 118 | 16 % | 20 534 | **41 %** | 1 | 93 % |
| *r_has_part* | fr | 178 286 | 40 855 | 23 % | 26 555 | **65 %** | 2 | 69 % |
| | en | 39 086 | 10 628 | 27 % | 3978 | **37 %** | 1 | 78 % |
| *r_matter* | fr | 16 419 | 4 547 | 28 % | 3728 | **82 %** | 1 | 87 % |
| | en | 1 709 | 575 | 34 % | 490 | **85 %** | 1 | 94 % |
| *r_object* | fr | 14 655 | 54 517 | 372 % | 10 592 | **19 %** | 1 | 94 % |
| | en | 7 088 | 9 190 | 129 % | 7 566 | **91 %** | 2 | 89 % |
| *r_carac* | fr | 32 585 | 58 809 | 180 % | 7 057 | **12 %** | 2 | 75 % |
| | en | 5 474 | 2 436 | 45 % | 2 094 | **86 %** | 1 | 94 % |
| *r_manner* | fr | 1 873 | 1 423 | 76 % | 1 109 | **78 %** | 2 | 77 % |
| | en | 1 751 | 5 938 | 339% | 1603 | **27 %** | 1 | 89 % |
| *r_location* | fr | 2 499 | 1 821 | 73 % | 1 071 | **59 %** | 1 | 89 % |
| *total* | - | 764 286 | 265 236 | - | 100 263 | - | - | - |
| *moyenne (arith.)* | - | - | - | 104 % | - | **57 %** | - | 80 % |

TABLE 1 – Inférence ascendante des relations sémantiques. Inférence ascendante des relations sémantiques **en→pivot** reflète l'intégration massive des relations taxonomiques depuis les ressources expertes telles que WordNet, RWN (Loukachevitch *et al.*, 2016).

L'expérience d'inférence descendante a concerné les partitions les moins peuplées, **es** et **ru**. La table 2 liste les résultats pour les relations principales de ces partitions. Nous faisons état du nombre de relations d'un type donné dans la partition lexicalisée avant inférence descendante (**avant_inf**), le nombre de nouvelles relations obtenues par inférence (**inf**) et évolution. L'état d'évolution du sous-graphe « ru » et « es » permet de justifier la démarche proposée.

Les résultats obtenus pour l'inférence ascendante montrent que la productivité d'inférence dépend de celle des autres processus de peuplement du $RLSM_{PI}$ tels que l'intégration des relations sémantiques

| type | l | #bef | #inf | #aft | ev |
|---|---|---|---|---|---|
| *r_isa* | ru | 46 827 | 7 036 | 53 863 | +14 % |
| | es | 36 807 | 268 040 | 304 847 | +828 % |
| *r_has_p.* | ru | 65 772 | 3 682 | 69 454 | +5 % |
| | es | 10 166 | 56 883 | 67 049 | +559 % |
| *r_mat.* | ru | 5190 | 4230 | 9 420 | +81 % |
| | es | 4013 | 7 351 | 7 764 | 183 % |
| *r_man.* | ru | 1 265 | 1 655 | 2 920 | +131 % |
| | es | 1 753 | 9 507 | 11 260 | +542 % |
| *r_loc.* | ru | 640 | 621 | 1 261 | +97 % |
| | es | 90 | 567 | 657 | +630 % |
| *by lang.* | ru | 119 694 | 17 224 | 136 918 | +14 % |
| | es | 52 739 | 342 348 | 395 087 | +649 % |
| *Totaux (moyenne)* | - | 172 433 | 359 572 | 532 005 | **+208 %** |

TABLE 2 – Descending inference of semantic relations. According to the size of the available corpora and external resources, the positive inference impact may vary.

à partir des ressources structurées. La précision obtenue par filtrage logique montre l'efficacité de ces filtrages dans le contexte multilingue. Cependant, la question de complexité peut être évoquée. Les prémisses d'une règle qui définissent le parcours à mettre en œuvre dans le cadre du filtrage logique sont destinées à réduire le nombre de relations à tester (et la complexité) en ne parcourant que certains types de relations. Cette opération de recherche des relations doit être exécutée $m$ fois. Ainsi, la complexité globale de filtrage logique (sans spécification de type de filtrage) serait de $O(m \times n^2)$. La **complexité moyenne** correspondrait dans ce cas au degré moyen du $\text{RLSM}_{\text{PI}}$ : $d_{av} = 4 \Rightarrow O(16 \times m)$.

# 3 Raffinements, cas particulier d'inférence endogène

Dans un réseau lexico-sémantique, la relation de raffinement permet de modéliser les sens d'usage d'un terme polysémique. Les raffinements correspondent à des cliques maximales (calcul) ou à des contributions des joueurs (jeu d'acquisition lexicale). Ils peuvent être nommés (glosés) ou non. Ainsi, pour le terme *baguette*, nous avons le sens « pain » par opposition aux autres sens (« encadrement », « bâton », « baguette magique »). Le raffinement glosé correspondant à ce sens est *baguette>pain*. Ainsi, nous avons la structure suivante dans notre ressource : *baguette* $\xrightarrow{r\_refinement}$ *baguette>pain* $\xrightarrow{r\_glose}$ *pain*. Un raffinement glosé peut être raffiné ou raffiné et glosé à son tour. Dans le cas d'une ressource qui possède déjà des relations de raffinement, il est possible de se baser sur les raffinements existants dans une de ses partitions pour inférer les sens dans les partitions qui ne contiennent pas ce type de relation. L'inférence des raffinements glosés se déroule en deux temps. Dans le cadre du ***schéma ascendant*** tous les sens présents dans les partitions lexicalisées sous forme de raffinements glosés se retrouvent dans le pivot interlingue. Ce pivot peut alors être considéré comme union raisonnée des parties monolingues de la ressource. Ce processus mis en œuvre à partir des partitions « en » et « fr », a permis d'atteindre le taux de raffinement du pivot interlingue de 30%.

Lorsque pour un terme il n'existe pas de raffinement glosé interlingue, mais ce terme possède plusieurs termes couvrant, le ***schéma de raffinement descendant en absence de raffinement nommé*** peut être appliqué. On considère les différents termes couvrants comme termes potentiellement liés à la glose. Initialement, les sens sont provisoirement étiquetés (ex. : étiquettes des termes couvrants). Ensuite,

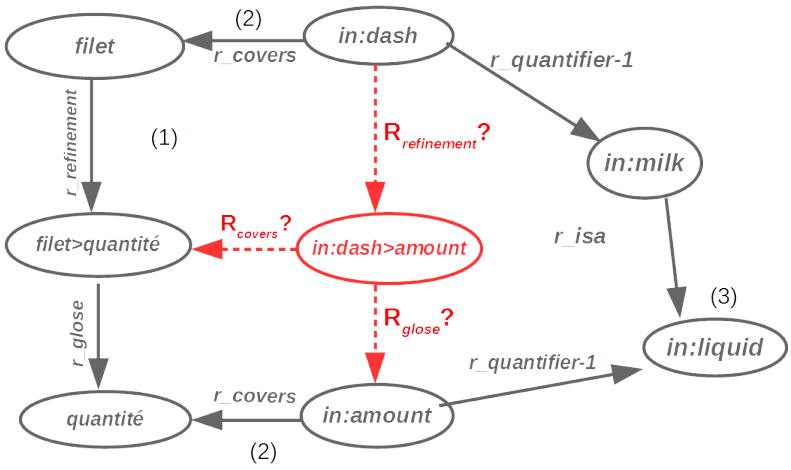



FIGURE 2 – Pour un terme raffiné (*filet*), son raffinement glosé (*filet>quantité*), la glose (*quantité*), le terme couvrant de terme raffiné (*in :dash*) et le terme couvrant de la glose (*in :amount*) : s'il existe une relation sémantique ou un chemin typé entre les termes couvrant le terme raffiné et la glose, un raffinement glosé interlingue peut être proposé. L'inférence implique la proposition de trois relations et la création d'un nouveau terme (raffinement interlingue).

| cible | fr>int | en>int | intersection | total |
|---|---|---|---|---|
| **int** | 8 558 | 33 930 | 254 | 31 752 |

TABLE 3 – Raffinements glosés interlingues acquis en appliquant le schéma ascendant.

on valide la distinction proposée en regroupant les sens redondants. L'étape finale du processus est le choix guidé par la sémantique de la glose appropriée dans la langue traitée. À ce jour ce schéma a permis de proposer 2 535 raffinements en russe tandis que 1 800 raffinements ont pu être obtenus pour cette langue grâce à l'inférence des raffinements glosés. Les mécanismes d'inférence proposés sont conçus pour améliorer le $\text{RLSM}_{\text{PI}}$ de façon continue.

# Conclusion

Dans le présent article, nous avons proposé une méthode d'inférence endogène des relations sémantiques. Nous avons ainsi exploité dans un cadre multilingue les méthodes conçues et précédemment déployées dans le cadre monolingue. Nous avons également proposé une méthode d'inférence des raffinements de sens à partir des raffinements existants et des contrastes observés dans la ressource. Cette méthode donne un rôle central au pivot évolutif amorcé en tant que pivot naturel et conçu pour devenir incrémentalement un pivot interlingue. Cette vision de pivot ainsi que l'inférence endogène des relations sémantiques confèrent à l'expérience que nous avons décrite un intérêt pour la construction des ressources qui impliquent des langues dites « peu dotées » pour lesquelles il n'existe que peu de ressources structurées et sémantiquement riches.

# Références

BOLLACKER K., EVANS C., PARITOSH P., STURGE T. & TAYLOR J. (2008). Freebase : a collaboratively created graph database for structuring human knowledge. In *In SIGMOD Conference*, p. 1247–1250.

CARLSON A., BETTERIDGE J., KISIEL B., SETTLES B., JR. E. R. H. & MITCHELL T. M. (2010). Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*.

CORDIER A., DUFOUR-LUSSIER V., LIEBER J., NAUER E., BADRA F., COJAN J., GAILLARD E., INFANTE-BLANCO L., MOLLI P., NAPOLI A. & SKAF-MOLLI H. (2014). *Taaable : A Case-Based System for Personalized Cooking*, In S. MONTANI & L. C. JAIN, Eds., *Successful Case-based Reasoning Applications-2*, p. 121–162. Springer Berlin Heidelberg : Berlin, Heidelberg.

DONG Z., DONG Q. & HAO C. (2010). Hownet and its computation of meaning. In *Proceedings of the 23rd International Conference on Computational Linguistics : Demonstrations*, COLING '10, p. 53–56, Stroudsburg, PA, USA : Association for Computational Linguistics.

FELLBAUM C. (1998). *WordNet An Electronic Lexical Database*. Cambridge, MA ; London : The MIT Press.

FERBER J. (1995). *Les systèmes multi-agents : vers une intelligence collective*. Paris : InterEditions.

GELBUKH A. F. (2018). Inferences for enrichment of collocation databases by means of semantic relations. *Computación y Sistemas*, **22**(1).

HERNÁNDEZ-GONZÁLEZ J., HRUSCHKA JR. E. R. & MITCHELL T. M. (2017). Merging knowledge bases in different languages. In *Proceedings of TextGraphs-11 : the Workshop on Graph-based Methods for Natural Language Processing*, p. 21–29, Vancouver, Canada : Association for Computational Linguistics.

HUA W., WANG Z., WANG H., ZHENG K. & ZHOU X. (2015). Short text understanding through lexical-semantic analysis. In *International Conference on Data Engineering (ICDE)*.

HUANG C.-R., TSENG I.-J. E. & TSAI D. B. (2002). Translating lexical semantic relations : The first step towards multilingual wordnets. In *COLING-02 : SEMANET : Building and Using Semantic Networks*.

JERMSURAWONG J. & HABASH N. (2015). Predicting the structure of cooking recipes. In L. MÀRQUEZ, C. CALLISON-BURCH, J. SU, D. PIGHIN & Y. MARTON, Eds., *EMNLP*, p. 781–786 : The Association for Computational Linguistics.

LAFOURCADE M. (2007). Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *SNLP'07 : 7th International Symposium on Natural Language Processing*, p.7̃, Pattaya, Chonburi, Thailand.

LAFOURCADE M. (2011). *Lexicon and semantic analysis of texts - structures, acquisition, computation and games with words*. Habilitation à diriger des recherches, Université Montpellier II - Sciences et Techniques du Languedoc.

LOUKACHEVITCH N., LASHEVICH G., GERASIMOVA A., IVANOV V. & DOBROV B. (2016). Creating russian wordnet by conversion. In *Dialog-2016*, p. 405–415, Moscow.

MÜLLER G. & BERGMANN R. (2015). Cookingcake : A framework for the adaptation of cooking recipes represented as workflows. In *Workshop Proceedings from The Twenty-Third International Conference on Case-Based Reasoning (ICCBR 2015), Frankfurt, Germany, September 28-30, 2015.*, p. 221–232.

RAMADIER L. (2016). *Indexation et apprentissage de termes et de relations à partir de comptes rendus de radiologie*. PhD thesis. Thèse de doctorat dirigée par Lafourcade, Mathieu Informatique Montpellier 2016.

SÉRASSET G. (2012). Dbnary : Wiktionary as a lmf based multilingual rdf network. In *LREC*.

SPEER R. & HAVASI C. (2012). Representing general relational knowledge in conceptnet 5.

TASSE D. & SMITH N. A. (2008). Sour cream :toward semantic processing of recipes. *T.R. CMU-LTI-08-005*, p.9̃.

# *Ma copie adore le vélo* : analyse des besoins réels en correction orthographique sur un corpus de dictées d'enfants

Jean-Yves Antoine[1], Marion Crochetet, Céline Arbizu, Emmanuelle Lopez[2], Samuel Pouplin[3], Amélie Besnier[4], Mathieu Thebaud[4]
(1) LIFAT, ICVL, 41000 Blois, France
(2) CRTLA, Hôpital R. Poincaré, 92380 Garches France
(3) PFNT, Hôpital R. Poincaré, 92380 Garches, France
(4) CMRRF de Kerpape, Ploemeur, 56270 France

Jean-yves.antoine@univ-tours.fr, marion.crochetet@aphp.fr, celine.arbizu@aphp.fr, emmanuelle.lopez@aphp.fr, abesnier@kerpape.mutualite56.fr, mthebaud@kerpape.mutualite56.fr

RÉSUMÉ

Cet article présente la constitution d'un corpus de textes produits, sur des données lors de dictées, par des enfants paralysés cérébraux (PC) ou dysorthographiques, son annotation en termes d'erreurs orthographiques, et enfin son analyse quantitative. Cette analyse de corpus a pour objectif de définir des besoins réels en matière de correction orthographique, et ce pour les personnes souffrant de troubles du langage écrit comme pour le grand public. Notre étude suggère que les correcteurs orthographiques ne répondent que partiellement à ces besoins.

ABSTRACT

**A corpus analysis to define the needs of dyslexic children in terms of spelling correction**

This paper presents a corpus of texts produced, during dictations, by dyslexic children, its annotation in terms of spelling errors, and finally its quantitative analysis. The purpose of this corpus analysis is to define real needs concerning automatic spelling correction, both for people suffering from language disorders and for ordinary people. Our study suggests that current spell checkers are unable to meet the majority of these needs.

MOTS-CLÉS : Correction orthographique, dyslexie, apprentissage des langues, analyse corpus

KEYWORDS: Spelling correction, dyslexia, language learning, corpus analysis

# 1 Introduction

La correction orthographique est une des applications les plus anciennes du TAL, puisque les premières recherches du domaine datent des années 1960s (Damereau 1964). Elle constitue une application emblématique du fait du rôle central de la maitrise d'une langue dans la représentation sociale d'un individu. Pourtant, si l'on en juge par sa faible représentation dans la littérature récente, la correction orthographique n'est plus considérée comme une problématique d'actualité. Est-ce à

dire que le verrou scientifique posé par la correction automatique est résolu? Nous pouvons en douter lorsque nous observons que les systèmes de correction automatique du commerce, mais également ceux issus de la recherche académique, peinent à détecter des erreurs orthographiques qui ne conduisent pas à la production d'un non-mot (cf § 4).

Dans cet article, nous abordons la question de la correction automatique dans un cadre particulier : celui des productions d'enfants apprenants en situation de handicap (paralysie cérébrale d'une part, dysorthographie d'autre part). La correction orthographique est destinée ici à être couplée à la prédiction de mots d'une système d'aide à la communication (Wandmacher et al. 2007). L'objectif n'est donc pas de corriger des énoncés complets, mais de prédire à la volée la suite d'une séquence de lettres qui est potentiellement erronée. Ainsi, si l'utilisateur a saisi le début de phrase *un ba..*, le système doit pouvoir prédire le mot *bateau*, mais également l'adjectif *beau,* pour corriger une éventuelle erreur d'encodage graphémique du phonème \o\. La combinaison correction/prédiction, explorée par (Li et al. 2013) est rarement envisagée, en particulier dans le cas de la dyslexie.

La tâche est donc plus complexe de celle envisagée pour les correcteurs orthographiques classiques. Pour mieux l'appréhender, nous avons mené une analyse des besoins sur des données issues de dictées réalisées par des enfants suivis au centre de rééducation de Kerpape (Mutualité du Morbihan) et à l'Hôpital Raymond Poincaré de Garches. L'annotation des erreurs orthographiques contenues dans le corpus, de même que l'analyse du comportement de correcteurs orthographiques sur cette ressource, nous a permis de dresser un ensemble de besoins qui sont autant de défis encore mal résolus. Cette étude reste préliminaire en termes de représentativité du corpus traité. Ses conclusions nous semblent toutefois assez éclairantes pour rouvrir la question d'une correction approfondie intégrée à la prédiction de mots.

Après un bref état de l'art sur la question, nous présentons le recueil et l'annotation du corpus sur lequel nous avons travaillé. Enfin, nous détaillons les résultats de l'analyse de besoins menée sur le corpus. En conclusion, nous esquissons une stratégie de correction qui sera mise en œuvre dans le cadre d'un projet, PREDICT4ALL, financé par la fondation Bennetot, fondation de la MATMUT.

# 2 Correction et erreurs orthographiques

## 2.1 Typologie des erreurs orthographiques

Les erreurs orthographiques qui surviennent dans un texte sont de nature variées. Plusieurs typologies d'erreurs ont été proposées dans la littérature, qui diffèrent par leurs objectifs. Certaines cherchent à rendre compte du comportement cognitif du scripteur et sont utiles à l'orthophoniste pour faire un bilan qui permettra l'adaptation du correcteur à son utilisateur. D'autres se concentrent uniquement sur la forme qu'aura à traiter le correcteur. (Kukich 1992) distingue ainsi :

- Les erreurs lexicales, qui conduisent à un non mot (mot absent du dictionnaire).
- Les erreurs syntaxiques qui conduisent à une phrase agrammaticale : par exemple, la partie du discours du mot n'est pas celle attendue, ou l'accord entre les mots n'est pas assuré.
- Les erreurs sémantiques, qui conduisent à une phrase incohérente sémantiquement.

Dans le cas des erreurs syntaxiques et sémantiques, le mot erroné mal orthographié correspond à une entrée du dictionnaire (*real-word errors*), ce qui rend leur détection plus délicate.

## 2.2 Correction orthographique : un bref état de l'art

**Correction hors contexte : distance d'édition** – Originellement, la correction automatique s'est focalisée sur les erreurs lexicales. Leur détection est immédiate, il reste à corriger le mot en cherchant dans le dictionnaire le mot le plus proche suivant un modèle d'erreur donné. Des modèles d'erreurs classiques sont la distance d'édition de Levensthein (1966), ou des modèles phonétiques comme Aspell (Atkinson 2011). La limite bien établie de ces approches est que les performances se dégradent lorsqu'on augmente la taille du dictionnaire: il y a de plus en plus de formes acceptables, ce qui fait qu'une erreur orthographique peut correspondre facilement à une autre forme lexicale : on se trouve alors dans le problème central de la détection des erreurs syntaxiques ou sémantiques.

**Correction contextuelle : ensemble de confusion** – Pour traiter les erreurs syntaxiques ou sémantiques, il est nécessaire de considérer le contexte d'occurrence du mot. Celui-ci n'est alors jugé correct que si sa probabilité d'apparition est supérieure à celle des mots qui lui sont proches orthographiquement, au sens du modèle d'erreur utilisé par l'approche hors-contexte. Chaque mot est ainsi associé à un ensemble de confusion (par exemple : *désert*, *dessert*) au sein duquel on cherche les termes de probabilité maximale suivant le modèle du canal bruité (Golding & Roth 1999, Norvig 2009). Le risque principal est de proposer une correction lorsque le mot est correct. Pour cette raison, les correcteurs privilégient la précision au rappel. Pour limiter la sur-correction, certains proposent de définir des listes noires de termes à ne pas corriger (Whitelaw et al. 2009), d'autres favorisent le mot saisi en n'envisageant son remplacement que si l'accroissement de probabilité dépasse un certain seuil (Jurafsky, Martin 2018). Le risque de sur-correction conduit également les correcteurs à définir les ensembles de confusion à une distance d'édition de 1 (erreurs simples). Ce choix se base sur les observations selon lesquelles 80% des erreurs orthographiques correspondent à une distance d'édition de 1 (Damereau 1964 ; Pollock et Zamora 1984). Un des enjeux de notre étude de corpus sera de savoir si cette conclusion faite sur l'anglais reste pertinente en français.

**Correction sémantique –**Enfin, certains auteurs envisagent une correction purement sémantique, en modélisant le contexte thématique d'apparition d'un mot par plongements de mots (Nagataa, 2017). Cette approche montre sa limite sur les langues à la morphologie suffisamment riche, pour lesquelles l'erreur peut être de flexion et ne pas conduire à un changement de lemme, donc de sémantique.

# 3 Corpus PREDICT4ALL d'erreurs orthographiques

L'objectif du projet PREDICT4ALL est de développer un moteur de prédiction de mots, utilisé dans un contexte d'aide à la communication orale ou écrite, qui s'adapte à des personnes en situation de handicap moteur et à des personnes présentant des troubles dysorthographiques ou simplement d'apprenants du français. Ces systèmes, tels le système Sibylle (Wandmacher et al. 2007), reposent sur une prédiction lexicale qui est perturbée par les erreurs orthographiques. Notre ambition est de combiner prédiction et correction à la volée. Les travaux sur la correction orthographique se sont rarement penchés sur le problème de l'apprentissage ou de la dyslexie. L'état de l'art récent se limite à notre connaissance à (Pedler 2007) et (Rello et al. 2015), qui ne portent pas sur le français.

Le corpus PREDICT4ALL réunit des données obtenues lors de dictées réalisées par 10 enfants scolarisés en CM (Cours Moyen de l'école primaire). A des fins de comparaison, les dictées étaient identiques et extraites de (Baneath 2015). Deux populations ont été distinguées :

– **DYS : Dyslexiques** – 5 enfants dysorthographiques (âge moyen 10 ans), diagnostiqués et suivis au centre de référence des troubles du langage de l'hôpital de Garches. Nous ne détaillons pas leur tableau clinique du fait de la taille réduite de la cohorte considérée.

– **PC** – 5 enfants paralysés cérébraux sans troubles langagiers, suivis au Centre Mutualiste de Rééducation et de Réadaptation Fonctionnelle de Kerpape. Ces apprenants souffrent de retards cognitifs qui ont un impact sur leur âge langagier. Plus qu'un groupe contrôle, ce public est la seconde cible visée par le projet PREDICT4ALL.

Ce corpus réunit 521 erreurs orthographiques. En dépit de sa taille limitée, seules deux autres ressources peuvent être comparées à la nôtre : d'une part un corpus anglais de 2654 erreurs mélangeant des sources assez hétérogènes (Pedler 2007) et d'autre part un corpus espagnol de 1171 erreurs (Rello and al. 2014). (Plisson et al. 2013) proposent un corpus de français canadien avec des enfants dysorthographiques en production libre et non pas en activité contrôlée de dictée. Ce corpus a l'avantage de la naturalité, mais ne peut permettre des comparaisons non biaisées entre scripteurs. En l'état actuel, la taille réduite de notre corpus interdit une caractérisation statistique du lien qui peut exister entre profil dyslexique et comportement langagier. Notre étude sera ainsi complétée, dans l'année à venir, sur une ressource de plus grande envergure en cours de constitution. L'étude pilote que nous présentons ici nous semble toutefois déjà présenter des résultats dignes d'intérêt.

Nous avons adopté un schéma d'annotation qui répond à deux besoins distincts : d'une part celui des chercheurs en TAL qui utilisent différentes ressources (lexiques, parseurs) pour analyser chaque chaîne de caractères. D'autre part, celui des orthophonistes qui étudient le comportement langagier de l'utilisateur, pour adapter au mieux la correction à son tableau clinique. Chaque erreur est caractérisée par les traits d'annotation suivants :

– **Erreur distincte** (TAL) – Précise si un mot comporte une ou plusieurs erreurs différentes. Plusieurs annotations distinctes pourront donc être distinguées dans un mot.
– **Type d'erreur** (TAL) – Lexicale / syntaxique / sémantique, inspirée de (Kukich 1992).
– **Morphologie** (TAL) **–** Impact de l'erreur en termes de forme, inspirée de (Damerau 1964). Nous distinguons d'une part les erreurs de segmentation : fusion (exemple : *lecole*) ou séparation de mots (exemple : *mon tagne*). D'autre part, en l'absence d'erreur de segmentation, nous comptabilisons la distance d'édition entre le mot écrit et le mot attendu.
– **Phonologie** (orthophonie) **–** Nous avons bâti une typologie basée sur la phonologie inspirée de (Catach 1980) tout en utilisant la terminologie erreurs phonologiquement plausibles (EPP) / non plausibles (ENPP) de (Martinet et Valdois, 1999). Celle-ci est utilisée en usage clinique par les orthophonistes pour repérer des compétences phonologiques déficitaires. Parmi les EPP, nous caractérisons les cas d'homophonie, le mauvais choix de graphème pour rendre un phonème (*le chamo*), les erreurs sur lettres muettes (*la souri*), les liaisons erronées (*dans sune des poches*), le mauvais encodage des semi-voyelles (*voillait*), les erreurs de flexion (*tu mange*) et celles portant sur des orthographes irrégulières (*fame* vs. *femme*). Pour les ENPP, nous distinguons les cas où l'erreur porte sur une graphie contextuelle, c'est-à-dire une graphie dont la prononciation dépend des voyelles environnantes (exemple : *gide* pour *guide*), les substitutions phonétiques, dont celles correspondant au changement d'un trait acoustique (voisement par exemple), les omissions ou insertions de graphèmes non muets, les erreurs séquentielles (*délcare* vs. *déclare*) et enfin les paragraphies morphémiques caractérisées, soit la production d'un mot qui partage avec le mot cible sa racine mais s'en différencie par un affixe (exemple : *ces*/*cette*, *un*/*une*...)

Enfin, nous avons étudié le comportement de plusieurs correcteurs grand public sur le corpus : le correcteur de *Microsoft Word 10* (2016), le correcteur en ligne *LanguageTool* (`languagetool.org/fr`) et enfin le correcteur *Cordial* de Synapse (version 11, 2005) qui présente l'intérêt de mettre en jeu une correction contextuelle avec analyse syntaxique profonde. Pour chaque erreur, nous distinguons les cas où (1) l'erreur se trouve parmi les propositions de correction, (2) l'erreur est détectée mais les propositions de correction sont erronées et enfin (3) l'erreur n'est pas détectée. La section suivante présente les analyses quantitatives réalisée sur cette ressource.

# 4 Analyse des besoins : étude quantitative du corpus

**Répartition des erreurs** – La distribution des erreurs permet de caractériser les besoins qui existent en matière de correction orthographique, pour les enfants PC et les enfants dysorthographiques.

| **Sous-corpus** | Nb. de mots | Nb. mots erronés | **Taux de mots erronés** | Nb. total d'erreurs | **Nb. d'erreur par mot erroné** | **% mots avec erreur multiple** |
|---|---|---|---|---|---|---|
| **PC** | 415 | 120 | **29 %** | 152 | **1,3** | **20 %** |
| **DYS** | 415 | 227 | **55 %** | 409 | **1,8** | **45 %** |

TABLE 1 : Distribution globale des erreurs dans le corpus PREDICT4ALL

La table 1 présente ces distributions sur les sous-corpus PC et DYS. Comme attendu, les troubles DYS rendent plus difficile la tâche de la correction. Nous observons comme (Plisson et al. 2013) une augmentation de la fréquence des erreurs chez les personnes DYS. Au final, plus de la moitié des mots (55%) du corpus DYS sont erronés. Ces erreurs sont par ailleurs plus profondes. Le taux moyen d'erreurs dans un mot erroné passe ainsi de 1,3 (PC) à 1,8 (DYS) : près de la moitié des mots erronés contiennent plusieurs erreurs chez les personnes DYS. Dans l'énoncé *il se défande contre les mousique*, les mots **défande* et **mousique* combinent une erreur de flexion avec respectivement un mauvais graphème ou une omission. Cette accumulation a un impact sur le type d'erreur.

**Type d'erreur** – Comme le montre la table 2, la proportion d'erreurs non lexicales se réduit ainsi de 49% à 29% entre les sous-corpus PC et DYS : l'accumulation des erreurs commises par les patients DYS réduit les chances de produire un mot du dictionnaire. Cette situation pourrait être un avantage pour la correction automatique, puisque ces erreurs sont détectables sans prise en compte du contexte. On observe toutefois que l'augmentation de la proportion d'erreurs lexicales est due pour partie à celle des erreurs de segmentation[1]. Or, la plupart de ces erreurs constituent un véritable défi pour la correction (par exemple : *janga jerer* pour de *j'engagerai*). Notons que ces problèmes de segmentation se retrouvent aussi chez les apprenants PC (9 % des cas). (Plisson et al. 2013) observe une même évolution (passage de 8% à 12% des cas) sur le français canadien.

| **Sous-corpus** | Erreurs syntaxiques | Erreurs sémantiques | **Total erreurs non lexicales** | Erreurs lexicales (hors segmentation) | *Erreurs de segmentation* |
|---|---|---|---|---|---|
| **PC** | 49% | 0% | **49%** | 42% | *9%* |
| **DYS** | 26% | 3% | **29%** | 56% | *15%* |

TABLE 2 : Distribution des erreurs du corpus PREDICT4ALL par type

Dans les deux populations, on observe enfin que le taux d'erreurs non lexicales est significativement supérieur à celui observé sur d'autres langues : 9% en espagnol chez (Rello et al. 2014) et 17% en anglais chez Pedler (2007). Une explication raisonnable à ce particularisme réside dans la forte morphologie flexionnelle du français. Ainsi, 47% des erreurs du sous-corpus PC concernent la flexion (table 4) et conduisent pour la plupart à des erreurs non lexicales (exemples : *explorer*, *explorai* et *exploré*) dont la détection nécessite une analyse syntaxique, même locale.

**Distance d'édition** – L'hypothèse selon laquelle ¾ des erreurs orthographiques sont à une distance d'édition de 1 (Damerau 1964) est retrouvée par (Rello et al. 2014) avec des personnes dyslexiques hispanophones (73% d'erreurs simples). Nos résultats ne confirment pas ces observations sur le

[1] Rares sont les erreurs de segmentation donnant des mots lexicaux, tel *ma copine a dore le vélo*

français (Table 3): les erreurs simples ne représentent que la moitié environ des situations dans nos deux populations. Dans près d'un quart des cas, la distance d'édition est même strictement supérieure à 2. Nos résultats recoupent ceux de (Pedler 2007) et (Baeza-Yates et Rello 2011) qui ne trouvent respectivement que 58% et 53% d'erreur simples sur l'anglais. Il s'agit d'un argument fort contre la définition d'ensembles de confusion limités à une distance d'édition de 1.

| **Sous-corpus** | Distance 1 | Distance 2 | Distance > 2 | **Total erreurs multiples** | *Erreur sur 1° caractères* |
|---|---|---|---|---|---|
| **PC** | 52 % | 26 % | 22 % | **48 %** | *4 %* |
| **DYS** | 46 % | 29 % | 25 % | **54 %** | *14 %* |

TABLE 3 : Distribution des erreurs en fonction de la position et de la distance d'édition

Enfin, notons que le taux d'erreurs portant sur la première lettre du mot varie entre 4 % (corpus PC) et 14 % (DYS). Ces ordres de valeur assez modestes se retrouvent sur l'anglais (Pedler 2007, Pollock et Zamora 1984, Yannakoudakis et Fawthrop 1983). Dans le cadre du projet PREDICT4ALL, la correction est couplée avec une prédiction lexicale. Le faible taux d'erreurs en début de mots nous dissuade de tenter une correction lorsque seule la première lettre du mot est déjà saisie. Trop de faux positifs seraient en effet prédits. Les corrections sur la première lettre seront envisagées uniquement lorsque 2 lettres auront été entrées par l'utilisateur.

**Phonologie** – La sévérité des erreurs observées chez les personnes DYS (erreurs de segmentation, distance d'édition élevée) incite à proposer une correction adaptée à l'utilisateur et non à envisager une correction générique. Cette adaptation s'appuiera sur le bilan, réalisé en orthophonie, des types d'erreurs systématiques de l'utilisateur. Nous ne détaillerons pas la répartition des types d'erreurs observées dans le corpus. La table 4 nous suggère toutefois que les troubles DYS augmentent le risque de produire des erreurs qui ne sont pas phonologiquement plausibles, observation déjà relevée par (Plisson et al. 2013), qui observe par ailleurs un même niveau général d'EPP.

| **Sous corpus** | **Phonologiquement plausible** | | | | **Phonologiquement non plausibles** | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Total** | *Graphème erroné* | *Flexion* | *Autre* | **Total** | *Substitut. phonétique* | *Insertion Omission* | *Séquence* | *Autre* |
| **PC** | **71 %** | *12 %* | *47 %* | *12 %* | **29 %** | *4 %* | *14 %* | *1 %* | *10 %* |
| **DYS** | **62 %** | *18 %* | *23 %* | *21 %* | **38 %** | *13 %* | *15 %* | *4 %* | *6 %* |

TABLE 4 : Distribution des erreurs en fonction de la position et de la distance d'édition

Les erreurs séquentielles (exemple : **graçon*) et les substitutions phonétiques (**crande* vs *grande*), qui sont marginales sur le corpus PC, représentent 17 % des erreurs des personnes DYS. Leur fréquence est toutefois variable d'une personne à l'autre et pourrait dépendre du type de dysorthographie. Le recensement de ces erreurs dans le corpus nous a toutefois permis d'obtenir des patrons de correction systématiques. Il nous semble ainsi important de définir, comme modèle d'erreur, des distances d'édition adaptées au tableau clinique de chaque personne.

# 5 Analyse des besoins : étude de performances

La correction que nous envisageons est réalisée à la volée en cours de saisie, ce que ne font pas les correcteurs standards. Nous avons toutefois examiné leur efficacité sur nos écrits d'enfants PC ou

dyslexiques. Les résultats (Table 5) mettent en évidence certaines limites des correcteurs, qui sont dans l'incapacité de corriger plus de la moitié des erreurs du sous-corpus PC. Dans plus d'un tiers des cas, ils ne détectent même pas ces erreurs, avant tout lorsqu'il s'agit d'erreurs non lexicales. Les taux de non-correction montent jusqu'aux trois quarts sur le sous-corpus DYS. Nos observations corroborent celles de (Pedler 2007) menées sur des correcteurs de l'anglais. On remarque enfin que mieux un correcteur se comporte sur le corpus PC, moins il s'en sort sur le corpus DYS. De fait, alors que *Cordial* corrige des erreurs syntaxiques négligées par les autres correcteurs sur le corpus PC, son parseur semble fortement désorienté sur les écrits DYS.

| **Corpus** | **PC** | | | **DYS** | | |
|---|---|---|---|---|---|---|
| **Erreur** | Corrigé | Détecté | Non détecté | Corrigé | Détecté | Non détecté |
| **LanguageTool** | **36 %** | 28 % | 36 % | **26 %** | 43 % | 31 % |
| **Word** | **42 %** | 19 % | 39 % | **24 %** | 40 % | 36 % |
| **Cordial** | **44 %** | 19 % | 37 % | **20 %** | 39 % | 41 % |

TABLE 5 : Performances des correcteurs orthographiques sur le corpus

# 6 Conclusion

Notre étude, originale sur le français, pose ou retrouve plusieurs recommandations pour une correction adaptée à des enfants présentant ou non une dysorthographie :

- La correction doit être envisagée à des distances d'édition de 2 ou 3 (et non 1),
- La correction sera adaptée à l'utilisateur, après bilan en orthophonie : on évitera les distances d'édition génériques au profit de distances paramétrables relevant de types d'erreurs précis,
- La correction ne doit pas être envisagée dès la saisie de la première lettre d'un mot,
- Les erreurs de segmentation ne peuvent être ignorées,
- Une correction contextuelle est essentielle au vu de l'importance des erreurs non lexicales.

Partant de cette analyse, nous avons développé un prototype de correction qui est intégré à la prédiction du système SIBYLLE. La prédiction est lancée à la fois sur les séquences saisies par l'utilisateur, mais également sur les chaînes corrigées respectant les contraintes ci-dessus. L'évaluation d'un premier prototype sur un extrait du corpus *Le Monde* montre que le taux d'économie de saisie de la prédiction (KSR-5) ne décroit que légèrement (47% contre 54% sans correction). La correction ne génère donc pas trop de faux positifs sur les énoncés corrects. Ses capacités de correction ont été testées sur le corpus d'erreurs WiCoPaCo (Max et Wisniewski 2010) issu de pages de révision de Wikipedia. Alors que la prédiction permet de corriger 69% des mots du corpus (dès qu'une bonne prédiction est proposée en cours de saisie, on choisit cette solution, évitant de fait certaines erreurs à venir), l'ajout de la correction élève ce taux à 74%. Ces premiers résultats ont été obtenus sur un prototype encore non optimisé. Ils sont toutefois suffisamment encourageants pour suggérer la pertinence d'une combinaison entre correction et prédiction dans le cadre de l'aide à la communication dédiée à des enfants PC ou dyslexiques.

# Remerciements

Cette recherche a été financée par la Fondation Bennetot, Fondation de la Matmut, dans le cadre du projet PREDICT4ALL.

# Références

Atkinson K. (2011). Gnu aspell.

Baeza-Yates R., Rello L. (2011) Estimating dyslexia in the web. Proc. Int. Cross-Displinary Conference on Web Accessibility, W4A'2011. Vol. 8, 1-4.

Baneath B., Alberti C., Boutard C., Gatignol P. (2015). *Chronodictées, outils d'évaluation des performances orthographiques*. Ortho-Editions ;

Catach N. (1980) *L'orthographe française: traité théorique et pratique avec des travaux d'application et leurs corrigés*. Coll. Nathan Université, Paris, Nathan.

Damerau F. (1964). A technique for computer detection and correction of spelling errors. *Communications of the A.C.M.*, 7:171–176

Golding A.R., Roth D. (1999). A Winnow based approach to context-sensitive spelling correction. *Machine Learning*, 34(1-3), 107–130

Jurafsky D., Martin J.H. (2018). Spelling erros correction and the noisy channel. In Jurafsky D., Martin J.H. *Speech & Language Processing* (3rd. ed.), Prentice Hall, Pearsons Ed. Apprendix B.

Kukich K. (1992). Techniques for automatically correcting words in text. *ACM Computing Survey*s, 24(4), 377–439

Lefavrais P. (2006). *Test de leximétrie de l'Alouette*. Paris: ECPA

Levensthein V. (1966) Binary codes capable of correcting deletions, insertions and reversions. *Soviet Physics Dokkady* 10, 845-848 (trad.)

Li A.Q., Sbattella L., Tedesco R. (2013) PoliSpell : an adaptive spellchecker and predictor for people with dyslexia. Proc. UMAP 2013. In Carberry S. et al. (Eds.) LNCS 7899, 302-309.

Martinet C., Valdois S. (1999). L'apprentissage de l'orthographe d'usage et ses troubles dans la dyslexie développementale de surface. *L'Année psychologique*, 99(4), 577-622

Max A., and Wisniewski G. (2010). Mining Naturally-occurring Corrections and Paraphrases from Wikipedia's Revision History, Actes *LREC 2010*. La Valette, Malte.

Nagataa R., Takamurab H., Neubigc G. (2017). Adaptive spelling error correction models for learner English. Actes de *KES'2017*, Marseille.

Norvig P. (2009). Natural language corpus data. In Segaran, T. and Hammerbacher, J. (Eds.), *Beautiful data: the stories behind elegant data solutions*. O'Reilly.

Pedler J. (2007). *Computer Correction of Real-word Spelling Errors in Dyslexic Texts*. PhD Thesis. London University.

Plisson A., Daigle D., Montésinos-Gelet I. (2013). The spelling skills of French-speaking dyslexic children. *Dyslexia*, 1 :76-91.

Pollock J.J., Zamora A. (1984). Automatic Spelling Correction in Scientific and Scholarly Text. *Communications of the A.C.M*. 27 (4): 358-68.

Rello R., Baeza-Yates R., Llisterri J. (2014). DysList: An Annotated Resource of Dyslexic Errors. Actes *LREC'2014*. 1289-1296.

Rello R., Ballesteros M., Bigham J.P. (2015). A Spellchecker for Dyslexia. Actes *ACM TASSETS'2015*.

Wandmacher T., Antoine J.-Y., Poirier F. (2007) Sibylle : a system for alternative communication adapting to the context and its user. Actes *ACM Conference on Assistive Technologies. ASSETS'2007*, Phoenix. Arizona. 203-210.

Whitelaw C., Hutchinson B., Chung G.Y., Ellis G. (2009). Using the web for language independent spellchecking and autocorrection. Actes de *EMNLP'2009*, 890–899.

Yannakoudakis, E.J., Fawthrop,D. (1983) The Rules of Spelling Errors. *Information Processing and Management.* 19 (2), 87-99.

# Modèles de langue appliqués aux schémas Winograd français

Olga Seminck Vincent Segonne Pascal Amsili
Laboratoire de Linguistique Formelle (Université Paris Diderot & CNRS)
8 place Paul Ricœur, 75013 Paris, Fance
olga.seminck@cri-paris.org, vincent.segonne@gmail.com,
pascal.amsili@linguist.univ-paris-diderot.fr

RÉSUMÉ

Les schémas Winograd sont des problèmes de résolution d'anaphores conçus pour nécessiter un raisonnement sur des connaissances du monde. Par construction, ils sont insensibles à des statistiques simples (co-occurrences en corpus). Pourtant, aujourd'hui, les systèmes état de l'art pour l'anglais se basent sur des modèles de langue pour résoudre les schémas (Trinh & Le, 2018). Nous présentons dans cet article une étude visant à tester des modèles similaires sur les schémas en français. Cela nous conduit à revenir sur les métriques d'évaluation utilisées dans la communauté pour les schémas Winograd. Les performances que nous obtenons, surtout comparées à celles de Amsili & Seminck (2017b), suggèrent que l'approche par modèle de langue des schémas Winograd reste limitée, sans doute en partie à cause du fait que les modèles de langue encodent très difficilement le genre de raisonnement nécessaire à la résolution des schémas Winograd.

ABSTRACT

**Language Models applied to French Winograd Schemas**

Winograd schemas are anaphora resolution problems built in such a way that reasoning about world knowledge is necessary to solve them. As a consequence, they are meant to be insensitive to simple statistical methods (based on coocurrence in corpus). However, today's state of the art systems use language models to solve Winograd schemas (Trinh & Le, 2018). In this paper, we report a study that tests similar language models on French Winograd schemas. It leads us to reconsider the evaluation metrics commonly used in the community. The results that we obtain, especially when compared to those of Amsili & Seminck (2017b), suggest that language model approaches to Winograd schemas remain limited, probably because language models, no matter how powerful they are, have difficulties to encode the kind of reasoning that is necesssary to solve Winograd schemas.

MOTS-CLÉS : Schémas Winograd, résolution d'anaphores, modèle de langue, context2vec, LSTM.

KEYWORDS: Winograd schemas, Anaphora Resolution, language model, Context2Vec, LSTM.

## 1 Introduction

Les schémas Winograd, proposés par Levesque *et al.* (2012) comme un test d'intelligence artificielle, sont des problèmes de résolution anaphorique conçus pour nécessiter un raisonnement sur des connaissances du monde. Chaque schéma est constitué d'une paire d'*items* : deux discours identiques à un mot (ou une expression) près, et qui comprennent une expression anaphorique à résoudre, dont l'antécédent change d'une version à l'autre. Ainsi, dans (1), la réponse naturelle pour la question

formée avec le mot faible (mot *spécial*) est *Nicolas* (R0), alors que la question formée avec le mot lourd (mot *alternant*) appelle la réponse *son fils* (R1).

(1) Nicolas n'a pas pu soulever son fils car **il** était trop ⟨faible/lourd⟩.
Qui était trop ⟨faible/lourd⟩ ? R0 : Nicolas
R1 : son fils

La collection de schémas en anglais publiée par Levesque *et al.* (2012) a fait l'objet de traductions dans d'autres langues : douze schémas ont été traduits en chinois, la collection entière a été traduite en japonais [1] et il existe une version de la collection traduite/adaptée en français à propos de laquelle il a été vérifié de plus que les schémas étaient 'Google-proofs' (non sensibles à des statistiques simples de co-occurrence de mots) et que les humains n'avaient pas de difficulté à les résoudre (Amsili & Seminck, 2017a,b).

Différentes approches ont été proposées pour résoudre les schémas Winograd, mais uniquement pour l'anglais. Nous présentons ici un premier système pour le français inspiré de méthodes à base de modèles de langue qui représentent actuellement l'état de l'art pour l'anglais (Trinh & Le, 2018) (exactitude de 63,7% sur 273 items Winograd). Ce résultat état de l'art est obtenu en combinant les scores de 14 modèles de langue distincts (qui utilisent différents algorithmes et corpus d'entraînement). N'ayant pas la possibilité d'en faire autant, nous présentons ici les résultats de deux modèles de langue : un modèle de langue récurrent et un modèle de contexte.

# 2 Méthode

## 2.1 Modèle de langue récurrent

Notre méthode est inspirée du travail de Trinh & Le (2018) ; elle consiste à générer, à partir de chaque item, une paire de phrases en remplaçant le pronom anaphorique par les antécédents possibles (2). Pour le français, ce remplacement demande parfois de ré-agencer la phrase, par exemple au moment de remplacer un pronom objet clitique (donc devant le verbe) par un groupe nominal plein (après le verbe). Nous avons toujours fait en sorte de ne pas présenter au modèle de langue des phrases agrammaticales. Après l'insertion des réponses, le modèle de langue permet d'attribuer une probabilité jointe à chaque phrase, l'hypothèse étant que celle qui obtient le meilleur score est celle qui contient la bonne réponse.

(2) a. Nicolas n'a pas pu soulever son fils car Nicolas était trop faible. (spe, R0)
b. Nicolas n'a pas pu soulever son fils car son fils était trop faible. (spe, R1)

Pour réaliser nos expériences, nous avons entraîné un modèle *Long Short Term Memory* (LSTM) (Hochreiter & Schmidhuber, 1997) sur la version française de Wikipedia [2]. Nous avons sélectionné les 100K mots les plus fréquents trouvés dans le corpus d'entraînement comme vocabulaire. Les

1. Les schémas chinois et japonais n'ont pas de publication associée. On peut les trouver sur la page : `https://cs.nyu.edu/davise/papers/WinogradSchemas/WS.html`. La traduction en chinois été faite par Wei Xu et la traduction en japonais par Soichiro Tanaka, Rafal Rzepka, et Shiho Katajima.
2. Nous avons utilisé la version fr-wikipedia 2016-12-12 construite par Coavoux (2017) avec l'outil de Giuseppe Attardi : `http://medialab.di.unipi.it/wiki/Wikipedia_Extractor`.

tailles des représentations vectorielles de mots et des couches cachées du réseau sont respectivement 1024 et 2048 et la minimisation de la perte a été réalisée avec l'algorithme *Adagrad* (Duchi *et al.*, 2011) avec un pas d'apprentissage de 0,2. Enfin un dropout de 0,25 a été appliqué sur la couche de sortie du LSTM.

Pour la résolution des schémas Winograd, l'application de cette méthode rencontre plusieurs types de problèmes et ne permet pas toujours au système de trancher entre les deux réponses possibles. Pour commencer, certains schémas sont atypiques dans le sens où il n'y a pas d'expression anaphorique que l'on puisse remplacer par les deux réponses possibles, voir par exemple (3) :

(3) Joël a vendu sa maison et en a acheté une nouvelle à quelques kilomètres.
Il va ⟨déménager/emménager⟩ ce jeudi.
Joël va ⟨déménager de/emménager dans⟩ quelle maison ?
R0 : son ancienne maison
R1 : sa nouvelle maison

Un autre problème vient du fait que les items Winograd peuvent être des discours constitués de plusieurs phrases (4), or le modèle de langue que nous utilisons n'est entraîné que sur des phrases isolées. On peut envisager plusieurs façons de gérer ce problème : soit concaténer les phrases, mais alors le modèle est confronté à des données assez différentes de ses données d'apprentissage, soit appliquer le modèle sur la dernière phrase, mais cette phrase seule ne contient en général pas les indices permettant la résolution, soit enfin entraîner des LSTM sur des discours de quelques phrases. Nous laissons de côté toutes ces solutions potentielles dans le présent travail, et décidons de ne pas répondre dans le cas d'un schéma multi-phrases.

(4) Fred est le seul homme encore vivant à se rappeler de mon arrière grand-père.
**C'** ⟨est/était⟩ un homme remarquable.
Qui ⟨est/était⟩ remarquable ?
R0 : Fred
R1 : mon arrière grand-père

Enfin, même si les modèles LSTM sont capables de gérer le cas de mots inconnus d'une phrase (en les remplaçant par un token unique ⟨UNK⟩), il peut arriver que les réponses (R0 et R1) d'un schéma soient justement inconnues du modèle : c'est le cas en particulier des noms propres. Or, ce sont les seuls mots qui distinguent les deux variantes, qui obtiendraient donc dans ce cas le même score, empêchant notre système de trancher. Notre système produit donc soit la réponse correspondant à la variante ayant le meilleur score, soit une non-réponse dans les cas évoqués ci-dessus.

## 2.2 Modèle contextuel

Context2vec (C2V) (Melamud *et al.*, 2016) est un modèle qui permet de représenter le contexte d'un token par le biais de réseaux de neurones récurrents bi-directionnels. Ces représentations permettent de mesurer à quel point un token est attendu dans un contexte. Nous utilisons ce modèle afin d'évaluer laquelle des deux réponses correspond le mieux au contexte. Ce modèle est donc différent des modèles LSTM utilisés par Trinh & Le (et de celui de la section 2.1) : au lieu de considérer la probabilité conjointe de la phrase comme nous l'avons vu pour les modèles LSTM, nous mesurons la similarité entre un mot cible et son contexte. Le modèle C2V construit des représentations pour le contexte et

pour le token cible dans le même espace vectoriel, les rendant donc comparables.

Comme pour l'expérience avec le modèle LSTM, on crée deux versions de chaque item en remplaçant l'anaphore par les réponses et en ré-agençant la phrase afin d'éviter des phrases agrammaticales. Ensuite, le token cible (en jaune) est sélectionné ; tous les autres tokens de la phrase forment le contexte (en gris). Pour la résolution des schémas, on calcule donc la similarité entre une réponse et son contexte, on choisit celle qui obtient le plus haut score.

(5) a. Simon a expliqué sa théorie à Marc, mais **Simon** ne l'a pas comprise.
b. Simon a expliqué sa théorie à Marc, mais **Marc** ne l'a pas comprise.

Les réponses des schémas Winograd se composent souvent de plusieurs tokens. Comme il n'est pas possible de sélectionner plus d'un token comme cible, il faut donc choisir quel token de la réponse devient la cible. Nous appliquons pour cela l'heuristique suivante : la tête syntaxique de la réponse est choisie comme cible, et le reste de la réponse fait partie du contexte.

(6) La table ne passe pas par la porte parce que **la** **table** est trop large.

Il y a cependant trois exceptions à cette règle : si la tête syntaxique de la réponse est un mot fonctionnel, nous prenons la tête en-dessous de la tête (par exemple, si une réponse est 'du garçon', nous prenons le mot 'garçon' comme cible). La deuxième exception intervient lorsque les têtes syntaxiques ne permettent pas de distinguer entre R0 et R1 comme dans l'exemple (3) où la tête de R0 et R1 est 'maison'. Dans ce cas, on prend les mots distinctifs ; pour l'exemple (3), on prendra donc 'ancienne' et 'nouvelle' comme cibles. La dernière exception est le cas où il y a plusieurs candidats pour être la tête syntaxique (e.g. la réponse *Dr. Vincenot*). Dans ce cas, nous prenons le premier token comme la tête, c'est-à-dire *Dr*.

Nous avons utilisé l'implémentation de Melamud *et al.* (2016) pour entraîner le modèle C2V. Cet entraînement a été réalisé sur le même corpus Wikipedia que le modèle LSTM présenté précédemment. En ce qui concerne les paramètres du modèle, nous avons fixé à 300 la taille des représentations vectorielles construites et nous avons utilisé l'algorithme d'optimisation *Adagrad* (Duchi *et al.*, 2011) avec un pas d'apprentissage de $0,001$. Nous avons considéré deux méthodes de calcul de similarité : la similarité cosinus et une similarité qui consiste à appliquer un log sigmoïde au produit scalaire des deux vecteurs que nous comparons. Dans les deux cas, plus la similarité est grande, plus le mot cible est compatible avec le contexte.

Le modèle C2V bute sur le cas des mots inconnus, quand il s'agit des mots cibles : on ne peut pas comparer le vecteur d'un mot inconnu à une représentation vectorielle de contexte. Or, il existe beaucoup de schémas Winograd dans lesquels les réponses R0 et R1 sont des noms propres ou d'autres mots inconnus, ce qui nous conduit à un taux de non-réponse important.

# 3 Évaluation

La tâche que nous voulons évaluer consiste à déterminer pour chaque item quel est l'antécédent parmi les deux candidats. Il y a deux réponses possibles (et une seule correcte), la mesure la plus naturelle

est donc l'*exactitude*, qui mesure la proportion de bonnes réponses parmi les items : $\frac{h}{n}$, où $h$ est le nombre de bonnes réponses et $n$ le nombre total d'items dans la collection.

Cette mesure ne permet cependant pas de distinguer les réponses incorrectes et les non-réponses. Pour prendre en compte la performance du système dans les seuls cas où il répond, on peut introduire une mesure que nous pourrions appeler *qualité* : $\frac{h}{n-\theta}$, où $\theta$ désigne le nombre de non-réponses. La qualité est une mesure de précision : c'est la proportion de bonnes réponses rapportée au nombre total de réponses. Mais il est important de noter que nous ne sommes pas dans un cas où, parmi un ensemble d'items, il s'agit de décider lesquels entrent dans une catégorie donnée. Autrement dit, nous ne sommes pas dans le cas où on définit habituellement, à partir d'une table de confusion, le couple de mesures *précision* et *rappel*[3]. Il nous semble par conséquent plutôt inapproprié d'appeler *rappel* le rapport $h/n$ (i.e. l'exactitude) comme le font Emami *et al.* (2018), et surtout de présenter la moyenne harmonique entre exactitude et qualité/précision (F1) comme une mesure synthétique pertinente. En effet ces deux mesures ne sont pas indépendantes : il n'est pas possible d'augmenter l'exactitude sans augmenter la qualité.

Dans le cas particulier des schémas Winograd, où il n'y a que deux réponses et où la collection est équilibrée par construction, un système répondant au hasard obtiendrait une exactitude de 50%. Ceci nous inspire une troisième mesure, que nous avons appelée *réussite* (Amsili & Seminck, 2017b) : c'est la mesure d'exactitude que donnerait un système qui répondrait au hasard pour tous les items correspondants à une non-réponse. La réussite s'interprète par comparaison avec le taux de 50% évoqué plus haut (niveau de la chance) : ce sont les points au-dessus de cette valeur qui nous renseignent sur la performance du système[4].

$$\textit{qualité} = \frac{h}{n-\theta} \qquad \text{exactitude} = \frac{h}{n} \qquad \text{réussite} = \frac{h+\frac{\theta}{2}}{n}$$

Nous présentons dans la table 1 les résultats de plusieurs systèmes, dont le nôtre, avec les différentes mesures évoquées à l'instant. Notons que si le système répond toujours, ce qui semble être le cas de celui de Trinh & Le (2018), $\theta = 0$ et toutes les mesures sont identiques. Les deux premières lignes de la table 1 présentent des résultats publiés par Emami *et al.* (2018) pour l'anglais. Le système AGQS, entièrement automatique, est présenté par ces derniers comme meilleur que le système de Sharma *et al.* (2015) sur la base de la mesure F1 qu'ils ont introduite (0,46 *vs.* 0,29). On peut critiquer cette conclusion, en observant que le système de Sharma *et al.* reste (légèrement) meilleur en terme de *réussite*. Il nous semble important de souligner de plus que le système MGQ de Emami *et al.*, cette fois-ci un système comprenant un traitement manuel, a certes de meilleurs résultats que AGQS quelle que soit la mesure, mais là encore il est contestable de poser qu'il dépasse celui de Sharma *et al.* (2015), du moins si on prend au sérieux la réussite telle que nous la proposons.

Nous donnons aussi dans la table 1 quelques-uns des résultats des travaux dont nous nous sommes directement inspirés (Trinh & Le, 2018) (toujours pour l'anglais). La seule mesure rapportée par les auteurs est l'exactitude, et nous en avons déduit que leur système est toujours capable de produire une réponse ($\theta = 0$). Cependant, peu de détails sont fournis (l'article est en cours d'évaluation). Si on

3. Situation dans laquelle la précision mesure à quel point le système est fiable lorsqu'il attribue la catégorie en question, et le rappel mesure à quel point le système est capable de retrouver tous les items appartenant à la catégorie. Il est bien établi que ces mesures sont indépendantes : pour augmenter la précision il faut réduire le nombre de *fausses alarmes* ($fa$), alors que pour augmenter le rappel il faut réduire le nombre de *manques* ($m$) ($h$ étant ici encore le nombre de bonnes réponses).

$$\text{précision} = \frac{h}{h+fa} \qquad \text{rappel} = \frac{h}{h+m}$$

4. Une autre façon, peut-être plus intuitive, de mesurer la performance pourrait consister à attribuer $+1/n$ aux bonnes réponses, $-1/n$ aux mauvaises, et 0 aux non-réponses. Soit $p$ cette mesure, qui va de $-1$ à 1 (elle vaut 0 si le système ne répond pas mieux que le hasard et 1 si le système répond toujours, et toujours correctement) ; on peut montrer qu'elle est définissable à partir de la réussite : $p = 2 \times \textit{réussite} - 1$.

| | $n$ | $h$ | $\theta$ | exactitude<br>*rappel**<br>$h/n$ | qualité<br>*précision**<br>$h/(n-\theta)$ | *F1** | réussite<br>$(h+\theta/2)/n$ |
|---|---|---|---|---|---|---|---|
| **Collection en anglais** | | | | | | | |
| Emami *et al.* (2018) AGQS | 273 | 106 | 83 | 38,83 | 55,79 | 45,79 | 54,03 |
| Emami *et al.* (2018) MGQ | 273 | 118 | 76 | 43,22 | 59,90 | 50,21 | 57,14 |
| Sharma *et al.* (2015) | 283 | 49 | 230 | 17,31 | 92,45 | 29,27 | 57,95 |
| Trinh & Le (2018) Word-full | 273 | 147 | 0 | 53,85 | | *—idem—* | |
| Trinh & Le (2018) 10 modèles | 273 | 168 | 0 | 61,54 | | *—idem—* | |
| Trinh & Le (2018) 14 modèles | 273 | 174 | 0 | **63,74** | | *—idem—* | |
| **Collection en français** | | | | | | | |
| LSTM (cet article) | 214 | 65 | 88 | 30,37 | 51,59 | 38,24 | 50,93 |
| C2V (cet article) | 214 | 33 | 158 | 15,42 | 58,93 | 24,44 | 52,34 |
| Amsili & Seminck (2017b) | 180 | 72 | 49 | 40,00 | 54,96 | 46,30 | 53,61 |

* terminologie de Emami *et al.* (2018)

TABLE 1 – Comparaison des métriques (exactitude, qualité, réussite) pour les collections anglaise et française. Rappelons que si le système répond toujours ($\theta = 0$) les trois mesures sont identiques.

peut supposer que, disposant de vocabulaires beaucoup plus important que nous, les auteurs n'ont pas rencontré de non-réponses dues à un défaut de vocabulaire, il serait utile de savoir comment a été traité le problème des schémas multi-phrases qui sont à l'origine de beaucoup de non-réponses dans notre cas.

La première ligne (Word-full) correspond au modèle qui nous a inspiré le plus directement : un modèle de langue LSTM basé sur les mots (*vs.* un modèle de caractères) dont les prédictions se font sur vocabulaire total de 800K mots avec des représentations vectorielles de mots de l'ordre de 1024 dimensions. Dans leurs différentes expériences, les auteurs ont utilisé plusieurs corpus d'entraînement pour les modèles de langue : LM-1Billion, CommonCrawl, SQuAD et Gutemberg books, mais ils ne précisent pas sur quel corpus le modèle Word-Full a été entraîné. Le système à 10 modèles consiste en un assemblage de plusieurs modèles, mélangeant modèles de mots et modèles de caractères avec des variations au niveau des paramètres des modèles (corpus d'entraînement, optimisation, profondeurs des réseaux, *etc.*). Tous ces modèles de langue ont cependant en commun le fait que la couche de sortie des réseaux produit un résultat dans le même espace vectoriel (1024 dimensions) permettant ainsi d'assembler conjointement les modèles. Enfin, le système avec 14 modèles reprend l'assemblage des 10 modèles précédents en ajoutant 4 nouveaux modèles entraînés spécialement pour la tâche : les auteurs ont construit des données d'entraînement spécifiques en extrayant les documents du corpus CommonCrawl ayant le plus de mots en commun avec les tokens trouvées dans les schémas Winograd. On peut noter que, quelle que soit la métrique utilisée, les systèmes combinant 10 et 14 modèles obtiennent des performances nettement meilleures que les autres système évoqués. On peut cependant aussi remarquer qu'un modèle standard, pas spécifiquement adapté à la tâche, n'atteint pas, seul, des ordres de grandeur meilleurs que les autres système état de l'art.

Nous présentons enfin les résultats obtenus par nos deux modèles de langue, qui obtiennent un résultat à peine meilleur que le hasard. Notons tout de même que le modèle C2V obtient de meilleurs scores que LSTM, même si ce dernier a beaucoup moins de non-réponses. Nous pouvons conclure qu'un modèle de langue non adapté à la tâche ne permet pas d'avoir un gain comparable à celui que Trinh & Le (2018) obtiennent en combinant 10 voire 14 modèles.

Nous avons reproduit en dernière ligne les résultats rapportés par Amsili & Seminck (2017b) qui ont utilisé un algorithme très simple basé sur l'information mutuelle entre les réponses et les mots spécial et alternant. Il est intéressant de noter que nos modèles de langue ne semblent pas avoir de meilleur résultat que leur système, qui était développé non pas pour résoudre les schémas, mais pour vérifier que les schémas français étaient « Google-proofs » (non sensibles aux statistiques de co-occurrence). La table montre que cet algorithme obtient quand-même 3,6 points au dessus du hasard, ce qui peut suggérer que la collection française n'est pas entièrement insensible aux statistiques de co-occurence. [5] Il est possible par conséquent que les gains de nos modèles soient essentiellement dûs à des schémas non « Google-proofs » (on peut noter qu'aucune vérification systématique de la *Google-proofness* n'a été faite pour la collection anglaise). Si c'est le cas, alors la meilleure performance de C2V peut s'expliquer par une capacité à mieux capter les informations de co-occurrence en ciblant les réponses, alors que le modèle LSTM construit une représentation globale de la phrase.

# 4 Conclusion

Notre étude suggère que les modèles de langue simples génériques ne permettent pas de résoudre les schémas Winograd, et que les quelques points gagnés par rapport au hasard sont au moins autant dûs au fait que certains schémas ne sont pas entièrement Google-proofs qu'à une capacité à encoder les informations pertinentes. Il semble cependant possible d'obtenir des résultats meilleurs avec des modèles de langue, mais seulement en les utilisant de façon très sophistiquée (et très coûteuse), de manière à parvenir à encoder des connaissances du monde ou des connaissances spécifiques à la tâche. Par ailleurs, si la méthode semble fonctionner sans grandes difficultés pour l'anglais (l'article de Trinh & Le (2018) ne donne pas tous les détails), nous avons constaté que le transfert au français a nécessité un important travail en partie manuel qui nous semble réduire la généralité de la méthode.

Nous pensons que la résolution des schémas Winograd se trouve devant une alternative : il est peut-être possible de faire appel à des modèles de langue encore plus sophistiqués, et de réfléchir à la façon dont de tels modèles peuvent encoder le genre de connaissance nécessaire pour ces schémas, mais on peut se demander si cette approche qui reste intrinsèquement statistique n'est pas en train d'atteindre une asymptote. L'autre option serait de renoncer aux modèles de langue pour basculer vers des approches plus spécifiquement orientées vers l'encodage du raisonnement, ce pour quoi ces schémas ont été explicitement conçus.

# Remerciements

Ce travail a reçu le soutien du *Labex* EFL (*Empirical Foundations of Linguistics*, ANR-10-LABX-0083). Nous remercions les relecteurs de TALN, ainsi que Chang Jiaqi, stagiaire du cursus de Linguistique Informatique à Paris Diderot.

5. Par exemple pour l'item *Un arbre est tombé sur le toit, il va falloir **le** déplacer* on peut s'attendre à des statistiques différentes de co-occurrence entre (*déplacer*, *toit*) et (*déplacer*, *arbre*).

# Références

AMSILI P. & SEMINCK O. (2017a). A Google-proof collection of French Winograd schemas. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017), co-located with EACL 2017*, p. 24–29.

AMSILI P. & SEMINCK O. (2017b). Schémas Winograd en français : une étude statistique et comportementale. In *Conférence sur le Traitement Automatique du Langage Naturel*, volume 2, p. 28–35, Orléans : Association pour le Traitement Automatique des Langues.

COAVOUX M. (2017). *Discontinuous Constituency Parsing of Morphologically Rich Languages*. PhD thesis, Univ Paris Diderot, Sorbonne Paris Cité.

DUCHI J., HAZAN E. & SINGER Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, **12**(Jul), 2121–2159.

EMAMI A., TRISCHLER A., SULEMAN K. & CHEUNG J. C. K. (2018). A generalized knowledge hunting framework for the Winograd schema challenge. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Student Research Workshop*, p. 25–31 : Association for Computational Linguistics.

HOCHREITER S. & SCHMIDHUBER J. (1997). Long short-term memory. *Neural computation*, **9**(8), 1735–1780.

LEVESQUE H., DAVIS E. & MORGENSTERN L. (2012). The Winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

MELAMUD O., GOLDBERGER J. & DAGAN I. (2016). context2vec : Learning generic context embedding with bidirectional LSTM. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, p. 51–61.

SHARMA A., VO N. H., ADITYA S. & BARAL C. (2015). Towards addressing the Winograd schema challenge-building and using a semantic parser and a knowledge hunting module. In *Proceedings of Twenty-Fourth International Joint Conference on Artificial Intelligence. AAAI*.

TRINH T. H. & LE Q. V. (2018). A Simple Method for Commonsense Reasoning. *ArXiv e-prints*.

# Multilingual and Multitarget Hate Speech Detection in Tweets

Patricia Chiril[1] Farah Benamara[1] Véronique Moriceau[1, 2]
Marlène Coulomb-Gully[3] Abhishek Kumar[4]
(1) IRIT, Université de Toulouse, France
(2) LIMSI, Univ. Paris Sud, Université Paris Saclay, France
(3) LERASS, Université de Toulouse, France
(4) Indian Institute of Science, India
patricia.chiril@irit.fr, benamara@irit.fr, moriceau@limsi.fr,
marlene.coulomb@univ-tlse2.fr, abhishekkumar12@iisc.ac.in

RÉSUMÉ

Les réseaux sociaux sont un espace où les utilisateurs sont libres d'exprimer leurs opinions ce qui donne lieu à la diffusion de messages haineux ou insultants qui doivent être modérés. Nous proposons dans cet article une approche supervisée pour la détection automatique de message haineux dans une perspective multilingue. Nous nous intéressons en particulier à la haine exprimée à l'encontre de deux types de cibles (des immigrants et des femmes) dans des tweets en anglais, ainsi qu'aux messages sexistes dans des tweets en anglais et en français. Divers modèles d'apprentissage automatique ont été développés, allant de modèles à base de traits, à des approches neuronales. Nos expérimentations montrent des résultats encourageants pour les deux langues.

ABSTRACT

Social media networks have become a space where users are free to relate their opinions and sentiments which may lead to a large spreading of hatred or abusive messages which have to be moderated. This paper proposes a supervised approach to hate speech detection from a multilingual perspective. We focus in particular on hateful messages towards two different targets (immigrants and women) in English tweets, as well as sexist messages in both English and French. Several models have been developed ranging from feature-engineering approaches to neural ones. Our experiments show very encouraging results on both languages.

MOTS-CLÉS : Réseaux sociaux, Détection de message haineux, Sexism, apprentissage supervisée.

KEYWORDS: Social media, Hate speech detection, Sexism, supervised learning.

## 1 Motivation

Social media networks such as Facebook, Twitter, blogs and forums, have become a space where users are free to relate events, personal experiences, but also opinions and sentiments about products, events or other people. This may lead to a large spreading of hatred or abusive messages which have to be moderated. In particular, theses messages may express threats, harassment, intimidation or "disparage a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic" (Nockleby, 2000). Although some countries, such as the United States, where hate speech is protected under the First Amendment as freedom of expression (Massaro, 1990), many other countries, such as France, have laws prohibiting it, laws that extend to the internet and social media. For instance, since the French law of 27 January 2017 related

to equality and citizenship, penalties due to discrimination are doubled (sexism is now considered as an aggravating factor). Gender equality has also been declared "major national cause" for the five-year period mandate of French president Emmanuel Macron [1]. In this context, it is important to automatically detect hateful messages on social platforms and possibly to prevent the widespreading of gender/racial stereotypes, especially towards young people.

From a computational point of view, hate speech detection is casted as a binary classification task : given a message, classify it as conveying a hateful content or not. Most studies focus on offensive contents in general while others on specific type of hate (like racism or hate speech against the LGBT community) relying on feature-based engineering or neural approaches (see (Schmidt & Wiegand, 2017)) for a comprehensive survey). Data are mainly tweets written in English, although some recent studies attempt to detect hate speech in Spanish (Anzovino *et al.*, 2018; Basile *et al.*, 2019), German (Ross *et al.*, 2017), Italian (Corazza *et al.*, 2018), Slovene (Fišer *et al.*, 2017) and Dutch (Jha & Mamidi, 2017), the latter focusing on benevolent sexist messages containing expressions such as *for a girl* or *like a man*. Other studies propose to tackle hate speech from a multilingual perspective (e.g., English and Spanish at IberEval2018 (Anzovino *et al.*, 2018)), but do not consider any cross-language experiments, as participants' models are trained and tested on each language separately. Finally, concerning French, hate speech detection only focuses on racist (Valette, 2004) or abusive messages (Papegnies *et al.*, 2017).

In this paper, we focus on (1) automatic hate speech detection towards *two different targets* – immigrants and women – and (2) automatic sexism detection *from a multilingual perspective*, namely in English and French tweets. For English, the data consists of tweets annotated as conveying hate speech against both immigrants and women, as part of HateEval@SemEval2019 (Basile *et al.*, 2019) (henceforth HS). For French, the data consists also of tweets, but annotated only for sexism (henceforth SEXISM). The main contributions of this paper are the following :

1. A new French dataset annotated for sexism detection.
2. A multitarget hate speech detection system. We propose both features-based models (relying on both language-dependent and language independent features) and a neural model to measure to what extent hate speech detection is target-dependent. When using the same model, our results show that HS achieve better results than SEXISM.
3. A multilingual hate speech detection. We also experiment with multilingual embeddings by training on one language and testing on the other in order to measure how the proposed models are language dependent. Our results are encouraging and open the door to hate speech detection in languages that lack annotated data for hate speech.

The paper is organized as follows. Section 2 presents the current state of the art, Section 3 describes our data, Section 4 the models and the experiments we carried out on multitarget detection while Section 5 on the multilingual experiments. We conclude providing some perspectives for future work.

# 2 Related work

Both sexism and racism can be expressed at different linguistic granularity levels going from lexical to discursive (Cameron, 1992) : e.g, women are often designated through their relationship with men or motherhood or by physical characteristics. Sexism can also be hostile or benevolent where messages are subjectively positive expressed in the form of a compliment (Glick & Fiske, 1996). Basically, sexism may be expressed explicitly or implicitly (see the following tweets from our French data) using different pragmatic devices, including :

1. http://www.egalite-femmes-hommes.gouv.fr/marlene-schiappa-presente-ses-priorites-

— Negative opinion, abusive message : *Meuf tu connais rien au foot. Tais toi. Contente de fan girler sur les joueurs et de mouiller sur MBappé ....*
— Stereotype : *C'est bon t'es une femme forte, te manque que la cuisine pour atteindre la perfection*
— Humor, irony : *Le fait maison c'est toujours mieux. La preuve, on préfère toujours sa femme à sa prostituée. #humour.*
— Benevolent sexism : *Elle court vite pour une femme*.

Same devices can also be employed towards immigrants, like the following tweet taken from the English data that illustrates a stereotype : *Illegals are dumping their kids heres o they can get welfare, aid and U.S School Ripping off U.S Taxpayers #SendThemBack! Stop Alowing illegals to Abuse the Taxpayer #Immigration*.

Most of the classifiers employed in hate speech detection still rely on supervised learning, and when creating a new classifier, one may manually design and encode different types of features from the data instances which will then be directly fed to the classical algorithms (Naive Bayes, Logistic Regression, Random Forest, SVM) or use deep learning methods that will automatically learn abstract features from data instances. Within the Automatic Misogyny Identification shared task at IberEval 2018, the best results were obtained with Support Vector Machine models with different feature configurations. There are also a few notable neural networks techniques deployed in order to detect hate speech in tweets that outperform the existing models : in (Badjatiya *et al.*, 2017) the authors used three methods (Convolutional Neural Network (CNN), Long short-term memory and FastText) combined with either random or GloVe word embeddings. In (Zhang & Luo, 2018) the authors implemented two deep neural network models (CNN + Gated Recurrent Unit layer and CNN + modified CNN layers for feature extraction) in order to classify social media text as racist, sexist, or non-hateful.

For most of the harassment and hate speech classification tasks, the most used information is depicted by the surface-level features (e.g. Bag of Words), the majority of authors choosing to include n-grams in the feature sets due to their high prediction rate. Due to the noise present in the data (especially on social media), many authors choose to combine the n-grams with a large section of additional features : linguistic features that take into consideration the POS information, dependency relations (long-distance relationship in between words), or word embeddings, which have the advantage of having similar vector representations for different, but semantically similar words. Since the task of hate speech detection and sentiment analysis are closely related, several approaches incorporate the latter as a supplementary classification step, assuming that generally negative sentiment relates to a hateful message (Dinakar *et al.*, 2012; Sood *et al.*, 2012).

Hate speech detection is a particularly difficult task mostly because in different contexts, the meaning of a message might change as it can be highly dependent on knowledge about the world. Because of this, in (Dinakar *et al.*, 2012), the authors present an approach in which they use automatic reasoning over aspects of the world. As it might be difficult to obtain knowledge about the world, the information about an utterance (meta information) may be used in order to refine unsatisfactory classification. For example, in (Waseem & Hovy, 2016), by using the users gender information the results were significantly improved, as the authors found that it is more likely for men to post hate speech messages [2]. This idea was further developed in (Hasanuzzaman *et al.*, 2017) where the authors introduced demographic aware information (age, gender and location) in order to tackle racism and confirm an important increase in performance. Another important feature is based on the

2. In spite of these findings and due to the difficulty of accurately identifying the gender of the user, we do not find this method favorable from an ethical perspective as we can encourage a gender bias in the system.

assumption that a user known for posting hateful messages is more likely to do so again in the future, thus by using the number of profane words in the users previous messages the detection performance improves (Dadvar *et al.*, 2013).

As far as we know, no work have addressed neither sexism detection in French, nor multitarget hate speech detection.

# 3 Data

Our data come from two corpora. The first one, HS-IW, is an already existing corpus containing English tweets annotated for hate speech against immigrants and women, as part of the HatEval task at SemEval2019. The second corpus, SEXISM, is new and contains French tweets collected between October 2017 and May 2018 with specific keywords such as *#balancetonporc, #sexisme*, names of politician women and men, insults, etc. The tweets have been labelled as sexist or non sexist by 3 annotators (2 female and 1 male annotators [3]). 329 tweets have been labelled by all annotators and the inter-annotator agreement is 0.89 (Cohen's Kappa). For these tweets, the final labels have been assigned according to a majority vote. Table 1 shows the distribution of the tweets for both tasks (hate speech and sexism detection).

| Task | #hate | #nonHate | Total |
|---|---|---|---|
| HS-IW (English) | 5,512 | 7,559 | 13,071 |
| SEXISM (French) | 659 | 2,426 | 3,085 |

TABLE 1 – Tweet distribution in both French and English datasets

# 4 Multitarget hate speech detection

Automatically labelling tweets as hateful/not hateful or sexist/not sexist is a challenging task because the language of tweets is full of grammatically and/or syntactic errors, it lacks conversational context, might consist of only one or a few words and because they can be indirectly hateful (through the use of sarcasm or irony) it makes the task of text-based feature extraction difficult. For both corpora, several models have been built, all tested using 10-cross-validation to better compare our results in cross-lingual experiments. In the next sections, we detail our models and then give our results.

## 4.1 Models

To measure to what extent hate speech detection is target-dependent, we propose several models raging from standard bag of words (our baseline), features-based models to neural model. For all the models, due to the noise in the data, we performed standard text pre-processing : removing user mentions, URLs, RT, stop words, degraded stop words and the words containing less than 3 characters were filtered out. For HS-IW, all the remaining words were stemmed using the Snowball Stemmer [4], while for SEXISM, tweets have been lemmatized using the French MSTParser [5]. We also experimented without stems and lemmas, but the results were not conclusive.

**Baseline.** In all experiments, we used as our baseline unigrams, bigrams and trigrams Tf/IDF (we ignored the terms that appear in less than 4 tweets, as well as the terms that appear in more than 80% of the tweets).

**Feature-based models.** We relied on state of the art features that have shown to be useful in hate speech detection. Our features include the following :

3. They are master degree's students in Communication and Gender.
4. `http://snowballstem.org`
5. `http://alpage.inria.fr/statgram/frdep/fr_stat_dep_mst.html`

— *Surface features :* such as the tweet length in words, the presence or absence of punctuation marks (sequence of question/exclamation marks), the presence of URLs and @user mentions.
— *Sentiment features* : The idea is to test whether identifying user's opinion can better classify his attitude as hateful or non-hateful. We took into consideration several existing lexicons : AFINN (Nielsen, 2011), SentiWordNet (Esuli & Sebastiani, 2006), Liu and Hu opinion lexicon [6], HurtLex (a multilingual hate word lexicon divided in 17 categories) (Bassignana *et al.*, 2018) and a lexicon containing 1 818 profanity English words created by combining a manually built offensive words list, the noswearing dictionary [7] and an offensive word list [8]. In the final models we chose to include only HurtLex and the lexicon we built, as none of the other models outperformed our baseline model. For the French corpus, we chose to use HurtLex, as it already contains hate words translated into French.
— *Emojis features* : We relied on a manually built emojis lexicon that contains 1 644 emojis along with their polarity among positive, negative and neutral.

We experiment with several combinations of the features above, and we finally keep the most relevant ones by applying the Chi2 feature selection algorithm. The best performing features have been used to train four classifiers ($C_1$, $C_2$ for the task of hate speech detection and $C_3$, $C_4$ for the task of sexism detection). For each classifier, we tried several machine learning algorithms (Naive Bayes, Logistic Regression, Support Vector Machine, Decision Tree and Random Forest) in order to evaluate and select the best performing one. Hereby, the hate speech baseline is a Random Forest (the number of trees in the forest = 360 with a maximum depth of the tree = 600) and the sexism baseline is a Support Vector Machine (linear kernel, C = 0.1). For $C_2$, best results have been obtained when using Random Forest only for intermediate classification, whose output were then combined and passed onto a final Extreme Gradient Booster classifier. The four classifiers are as follows :

— $C_1$ : combines the length of the tweet with the number of words in the profanity lexicon with a baseline architecture as described above
— $C_2$ : on top of $C_1$ features we also used the number of positive and negative emojis and emoticons and we perform linear dimensionality reduction by means of truncated Singular Value Decomposition (latent semantic analysis on TF/IDF matrices).
— $C_3$ : combines the length of the tweet with the number of words in the HurtLex lexicon on top of a baseline architecture
— $C_4$ : the same features as $C_3$ but with a $C_2$ system architecture

**Neural model.** The last model used a Bidirectional LSTM with an attention mechanism that attends over all hidden states and generates attention coefficients [9]. The hidden states were then averaged using the attention coefficients in order to generate the final state which was then fed to a one-layer feed-forward network for obtaining the final label prediction. For the task of hate speech detection, we used pre-trained on tweets [10] Glove embeddings with an embedding dimension of 200 (Pennington *et al.*, 2014), while for the task of sexism detection we used pre-trained on Wikipedia and Common Crawl FastText French word vectors with an embedding dimension of 300 (Grave *et al.*, 2018)). We experimented with different hidden state vector sizes, dropout values and attention vector sizes. The results reported in this paper were obtained by using 300 hidden units, an 150 attention vector, a dropout of 50% and the Adam optimizer with a learning rate of $10^{-3}$. For the BiLSTM we used a Relu activation function and we run all the experiments for maximum 100 epochs, with a patience of

6. https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon
7. https://www.noswearing.com/dictionary
8. http://www.cs.cmu.edu/~biglou/resources/bad-words.txt
9. We also experimented with other neural architectures, like CNN, but the results were lower.
10. We also experimented with pre-trained on Wikipedia word vectors, however the accuracy decreased by 3%

10 and batch size of 64 [11].

## 4.2 Results

Since the number of sexist instances in the French corpus is relatively small, the results presented in this paper were obtained by using 10-cross validation. Table 2 shows how the experiments were set up and presents the results in terms of accuracy (A), macro-averaged F-score (F), precision (P) and recall (R). The best results in terms of macro-averaged F-score (the evaluation metric used for ranking at SemEval) are presented in bold, while the columns left empty were intentionally left so, as we employed same system architectures with different features for the two tasks. Overall, our results show that when using the same model, the results achieved for the task of hate speech detection are better than the results for sexism detection.

| | **Hate speech detection** | | | | **Sexism detection** | | | |
|---|---|---|---|---|---|---|---|---|
| | A | F | P | R | A | F | P | R |
| Baseline | 0.772 | 0.762 | 0.764 | 0.669 | 0.827 | 0.676 | 0.734 | 0.335 |
| $C_1$ | **0.788** | **0.780** | **0.785** | **0.684** | – | – | – | – |
| $C_2$ | 0.781 | 0.778 | 0.754 | 0.723 | – | – | – | – |
| $C_3$ | – | – | – | – | 0.830 | 0.441 | 0.751 | 0.306 |
| $C_4$ | – | – | – | – | **0.822** | **0.688** | **0.665** | **0.386** |
| BiLSTM + attention | 0.736 | 0.727 | 0.709 | 0.646 | 0.77 | 0.497 | 0.416 | 0.522 |

TABLE 2 – Hate speech detection and sexism detection results in both HS and SEXISM corpora

Among the systems, $C_1$ represents our best performing one for the task of hate speech detection, while $C_4$ performed best for the task of sexism detection.

**Error analysis :** A manual error analysis of the instances for which our best performing model and manual annotation differ shows that in the misclassification of hateful instances intervene several factors : the presence of off-topic tweets, the lack of context (as some words that trigger hate in certain contexts may have different connotations in others) and implicit hate speech that employs stereotypes or metaphors in order to convey hatred. We also identified tweets for which we question the original label when taking into account the class definition. Below, we have provided some examples.

**Example 1 (HS-IW) :** Although in the first tweet (annotated as not hateful) the user talks about Donald Trump, which doesn't fit in the targeted categories (immigrants or women), the annotation raises problems when trying to classify tweets such as the second one (annotated as hateful).

— I love my religious brothers and sisters, but @realDonaldTrump, FUCK YOU, YOU'RE NOT EVEN A REAL THEOCRAT YOU FAT USLESS BITCH.
— @menzemerized_ Worse i have proof. A picture i took of you and one you took of me on the same night. Useless ungreatful kunt !

**Example 2 (SEXISM) :** Both of the following tweets were misclassified due to the lack of context and knowledge about the world. In the first tweet, as we don't have enough information about the "liberté d'importuner" movement, we aren't able to properly classify the disagreement of the user with Catherine Deneuve's statements. The same problem arises in the second tweet, as the speech employs irony.

— Ce que je pense de la "liberté d'importuner". #Sexisme #CatherineDeneuve #Tribune C'est pas parce que vous aimez la soumission qu'on doit toutes apprécier. L'avis des vieilles bourgeoises qui ne prennent plus le métro sur les frotteurs, on s'en passe.

11. The hyperparameters were tuned on the validation set (20% of the training dataset), such that the best validation error was produced.

— Merkel en Allemagne. Thatcher et maintenant #TheresaMay au Royaume-Uni. En France une femme présidente ? Folie ! Décadence !

# 5 Multilingual hate speech detection

We also experimented with multilingual embeddings : Glove bilingual word embeddings [12] obtained as described in (Ferreira *et al.*, 2016) as well as French and English FastText word vectors mapped into the same embedding space following the alignment approach presented in (Smith *et al.*, 2017). For the experiments we used the same BiLSTM model described in Section 4.1, firstly by using the HS-IW English corpus for training and the SEXISM French corpus for testing, and secondly by using jointly the two corpora (HS-IW and 30% of the original SEXISM corpus) for training and testing on the remaining SEXISM corpus. Table 3 shows how the experiments were set up and presents the results in terms of accuracy (A) and macro-averaged F-score (F), the best result in terms of accuracy being presented in bold.

| **Corpus** | | **FastText** | | **Glove** | |
|---|---|---|---|---|---|
| Train | Test | A | F | A | F |
| English | French | 0.783 | 0.445 | 0.732 | 0.485 |
| English + French | French | **0.790** | **0.461** | 0.766 | 0.479 |

TABLE 3 – Multilingual hate speech detection results

The multilingual experiments results are somewhat comparable to the results obtained when training and testing on the French data (cf. Table 2). This is very encouraging as one can rely on external annotated data for sexism in other languages to learn a model on a different language. Of course, these results have to be confirmed as for the moment we do not have the actual distribution of the tweets in the SemEval corpus (the number of tweets that convey hate towards immigrants and the number of tweets that convey hate towards women).

**Error analysis :** The error analysis shows that in the absence of context and knowledge about the world (the #balancetonporc movement, as well as the persons to which the author of the tweet is referring to) and without employing irony detection systems, we misclassify (as non-sexist) tweets such as the following one :

— Donc on va avoir une conférence à Sciences Po avec Raphaël Enthoven, Aurore Bergé, Elisabeth Lévy et Pierre-Oliver Sur pour se demander comment #balancetonporc favorise la délation et la "mise au pilori" des accusés wow so much progressisme et ouverture.

# 6 Conclusion

This paper proposed several models that can be used in order to identify messages that convey hate and proved the portability of these systems for the task of detecting sexist messages in French. As far as we know, this is the first work on sexism detection in French on Twitter data, this study serving as a first step towards improving the task. In our future work we plan on studying ways to retrieve contextual information, and as the results seemed promising, we also plan on experimenting more in a multilingual embedding space.

# Acknowledgments

This work has been funded by Maison des Sciences de l'Homme et de la Société de Toulouse under the project AMeSexTo.

12. http://www.cs.cmu.edu/~afm/projects/multilingual_embeddings.html

# Références

ANZOVINO M., FERSINI E. & ROSSO P. (2018). Automatic identification and classification of misogynistic language on twitter. In *International Conference on Applications of Natural Language to Information Systems*, p. 57–64 : Springer.

BADJATIYA P., GUPTA S., GUPTA M. & VARMA V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, p. 759–760 : International World Wide Web Conferences Steering Committee.

BASILE V., BOSCO C., FERSINI E., NOZZA D., PATTI V., RANGEL F., ROSSO P. & SANGUINETTI M. (2019). Semeval-2019 task 5 : Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)* : Association for Computational Linguistics", location = "Minneapolis, Minnesota.

BASSIGNANA E., BASILE V. & PATTI V. (2018). Hurtlex : A multilingual lexicon of words to hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018*, volume 2253, p. 1–6 : CEUR-WS.

CAMERON D. (1992). *Feminism and Linguistic Theory.* Palgrave Macmillan.

CORAZZA M., MENINI S., ARSLAN P., SPRUGNOLI R., CABRIO E., TONELLI S. & VILLATA S. (2018). Comparing Different Supervised Approaches to Hate Speech Detection. In *EVALITA 2018*, Turin, Italy.

DADVAR M., TRIESCHNIGG D., ORDELMAN R. & DE JONG F. (2013). Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*, p. 693–696 : Springer.

DINAKAR K., JONES B., HAVASI C., LIEBERMAN H. & PICARD R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, **2**(3), 18.

ESULI A. & SEBASTIANI F. (2006). Sentiwordnet : A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06*, p. 417–422.

FERREIRA D. C., MARTINS A. F. & ALMEIDA M. S. (2016). Jointly learning to embed and predict with multiple languages. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, volume 1, p. 2019–2028.

FIŠER D., ERJAVEC T. & LJUBEŠIĆ N. (2017). Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in slovene. In *Proceedings of the First Workshop on Abusive Language Online*, p. 46–51.

GLICK P. & FISKE S. T. (1996). The ambivalent sexism inventory : Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, **70**(3), 491–512.

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

HASANUZZAMAN M., DIAS G. & WAY A. (2017). Demographic word embeddings for racism detection on twitter. In *IJCNLP*.

JHA A. & MAMIDI R. (2017). When does a compliment become sexist ? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, p. 7–16.

MASSARO T. M. (1990). Equality and freedom of expression : The hate speech dilemma. *Wm. & Mary L. Rev.*, **32**, 211.

NIELSEN F. Å. (2011). A new ANEW : evaluation of a word list for sentiment analysis in microblogs. In M. ROWE, M. STANKOVIC, A.-S. DADZIE & M. HARDEY, Eds., *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts' : Big things come in small packages*, volume 718 of *CEUR Workshop Proceedings*, p. 93–98.

NOCKLEBY J. T. (2000). Hate speech. In *Encyclopedia of the American Constitution (2nd ed., edited by Leonard W. Levy, Kenneth L. Karst et al.*, p. 1277–1279.

PAPEGNIES E., LABATUT V., DUFOUR R. & LINARÈS G. (2017). Detection of abusive messages in an on-line community. In *14ème Conférence en Recherche d'Information et Applications (CORIA)*, p. 153–168.

PENNINGTON J., SOCHER R. & MANNING C. D. (2014). Glove : Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, p. 1532–1543.

ROSS B., RIST M., CARBONELL G., CABRERA B., KUROWSKY N. & WOJATZKI M. (2017). Measuring the reliability of hate speech annotations : The case of the european refugee crisis. In *Proceedings of NLP4CMC III : 3rd Workshop on Natural Language Processing for Computer-Mediated Communication (Bochum)*, p. 6–9.

SCHMIDT A. & WIEGAND M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, p. 1–10.

SMITH S. L., TURBAN D. H. P., HAMBLIN S. & HAMMERLA N. Y. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *CoRR*, **abs/1702.03859**.

SOOD S. O., CHURCHILL E. F. & ANTIN J. (2012). Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology*, **63**(2), 270–285.

VALETTE M. (2004). Sémantique interprétative appliquée à la détection automatique de documents racistes et xénophobes sur internet. In *Colloque International sur le Document Electronique*, p. 215–230 : Centre de recherche en Ingénierie Multilingue, INaLCO.

WASEEM Z. & HOVY D. (2016). Hateful symbols or hateful people ? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, p. 88–93.

ZHANG Z. & LUO L. (2018). Hate speech detection : A solved problem ? the challenging case of long tail on twitter. *arXiv preprint arXiv :1803.03662*.

# Observation de l'expérience client dans les restaurants

Iris Eshkol-Taravella, Hyun Jung KANG
UMR 7114 MoDyCo - CNRS, Université Paris Nanterre, France
ieshkolt@parisnanterre.fr, hyunjung.kang@parisnanterre.fr

RÉSUMÉ

Ces dernières années, les recherches sur la fouille d'opinions ou l'analyse des sentiments sont menées activement dans le domaine du Traitement Automatique des Langues (TAL). De nombreuses études scientifiques portent sur l'extraction automatique des opinions positives ou négatives et de leurs cibles. Ce travail propose d'identifier automatiquement une évaluation, exprimée explicitement ou implicitement par des internautes dans le corpus d'avis tiré du Web. Six catégories d'évaluation sont proposées : opinion positive, opinion négative, opinion mixte, intention, suggestion et description. La méthode utilisée est fondée sur l'apprentissage supervisé qui tient compte des caractéristiques linguistiques de chaque catégorie retenue. L'une des difficultés que nous avons rencontrée concerne le déséquilibre entre les classes d'évaluation créées, cependant, cet obstacle a pu être surmonté dans l'apprentissage grâce aux stratégies de sur-échantillonnage et aux stratégies algorithmiques.

ABSTRACT

**Mapping Reviewers' Experience in Restaurants**

In opinion mining and sentiment analysis, studies have focused on extracting either positive or negative opinions from text and determining the targets of these opinions. Beyond the opinion polarity and its target, we propose a corpus-based model that detects evaluative language at a finer-grained level. Moreover, whereas previous works assume that classes are evenly distributed, the classes were highly imbalanced in our work. We chose a dataset of online restaurant reviews in French. We used machine learning methods to detect and classify evaluative language. Furthermore, we used resampling and algorithmic approaches to deal with class imbalance problem.

MOTS-CLÉS : Fouille d'opinion, Langage évaluative, Données disparates, Expérience client.

KEYWORDS: Opinion mining, Evaluative language, Imbalanced data, Customer experience.

## 1 Introduction

Les avis en ligne aboutissent aujourd'hui à une production abondante de données véhiculant l'évaluation du consommateur vis-à-vis de son expérience. En TAL, c'est le domaine de la fouille d'opinions qui s'occupe de cette évaluation. Pour exprimer l'évaluation, plusieurs termes peuvent être utilisés : opinion, sentiment, attitude, affect, subjectivité, etc. (Benamara *et al.* 2017). Les travaux existants se distinguent selon trois axes : la fouille d'opinion au niveau du document (Pang *et al.*, 2002 ; Turney, 2002), au niveau de la phrase (Hatzivassiloglou & Wiebe, 2000 ; Riloff & Wiebe, 2003 ; Yu & Hatzivassiloglou 2003 ; Kim & Hovy, 2004 ; Wiebe *et al.*, 2004 ; Wilson *et al.*, 2004 ; Riloff *et al.*, 2006 ; Wilson *et al.*, 2006) ou au niveau de l'aspect (Hu & Liu, 2004 ; Liu, 2015). Les recherches se limitent majoritairement aux seules catégories d'opinions positives et négatives (Pang *et al.*, 2002 ;

Turney, 2002 ; Hu & Liu, 2004 ; Kim & Hovy, 2004 ; Pontiki *et al.*, 2014, 2015, 2016). Cependant, ces dernières années, les recherches se sont réorientées vers l'extraction de suggestions (Brun & Hagege, 2013 ; Negi *et al.* 2015, 2016, 2018). Ainsi l'atelier SemEval 2019 propose des tâches visant l'extraction des suggestions dans des forums et des avis sur les hôtels. Cette tendance montre que l'intérêt que les chercheurs ont pour l'évaluation sur le web dépasse largement la notion d'opinion positive ou négative. Ainsi Benamara *et al.* (2017) proposent la notion d'intention lors d'évaluation d'un produit.

Dans cet article, nous nous intéressons aussi à la détection automatique de l'évaluation. Nous considérons que l'évaluation est une notion complexe puisqu'elle ne se limite pas juste à une valeur positive ou négative. Nous proposons un modèle d'évaluation fondé sur l'observation manuelle du corpus des avis postés en ligne sur des restaurants. Ce modèle est composé de 4 catégories : l'*opinion* (*positive*, *négative*, *mixte*), la *suggestion*, l'*intention* et la *description* [1]. Les expériences de leur détection sont fondées sur un apprentissage supervisé qui tient compte des caractéristiques linguistiques de chaque catégorie. L'une des difficultés que nous avons rencontrée concerne le déséquilibre entre les classes d'évaluation créées, cependant, cet obstacle a pu être surmonté dans l'apprentissage grâce aux stratégies de sur-échantillonnage et aux stratégies algorithmiques.

# 2 Modélisation

L'objectif du travail présenté est de détecter l'évaluation d'un restaurant exprimée soit explicitement, soit de manière implicite par un consommateur en ligne. Notre modèle s'appuie sur quatre éléments que l'on détaillera plus bas : l'opinion (positive, négative, mixte), la suggestion, l'intention et la description.

**L'opinion** : il s'agit d'une notion générale largement utilisée dans la littérature. Elle concerne l'idée que le consommateur se fait du restaurant et comment il le qualifie, le lexique évaluatif constitue donc l'indice le plus important pour détecter l'opinion. Dans le corpus, les adjectifs comme « bon », « excellent », « parfait » ou « délicieux » sont généralement associés aux opinions positives, tandis que les termes « cher », « dommage », « déception » et « bruyant » possèdent une connotation négative. Dans cette étude, les opinions telles que « pas mal », « correct », « sans plus » ont été considérées comme des opinions positives de faible intensité. La polarité d'une phrase donnée peut cependant varier selon les éléments contextuels (aussi nommés modificateurs) qui infléchissent la valeur initiale d'un terme (la négation, les intensifieurs, les atténuateurs, les conjonctions). On parle de polarité mixte lorsque l'opinion comporte les deux polarités (positive et négative). Dans de nombreux cas, celles-ci coexistent et s'articulent autour de la conjonction « mais ». Plus précisément, une polarité s'y trouve inversée, comme dans les exemples suivants : « Plat très bon, mais dessert médiocre. », « Accueil très sympathique mais cuisine décevante ». Une analyse de notre corpus montre que 64% des opinions mixtes contiennent la conjonction « mais ». En outre, certains locuteurs peuvent donner une évaluation positive en montrant leur gratitude à travers les interjections comme « bravo », « merci », etc., en mentionnant le chef ou d'autres membres du personnel : « Bravo au chef ! », « Merci pour ce bon dîner. ».

**La suggestion** : c'est une expression d'un conseil émis par un consommateur. Les suggestions peuvent être adressées à la fois aux restaurants (afin qu'ils prennent conscience des problèmes) et aux autres clients potentiels (futurs consommateurs). Elles sont souvent exprimées par les verbes

1. Les classes seront détaillées dans la section 2.

dénotant l'action. Il s'agit d'une action qui doit être réalisée par le restaurant visité ou par les futurs visiteurs selon le conseil de l'auteur d'un avis. La suggestion se manifeste fréquemment à travers le pronom « vous » ou l'impératif de la deuxième personne du pluriel, ainsi qu'à travers le conditionnel. Par exemple : « **N'hésitez** pas à venir découvrir ce restaurant, **vous** ne le **regretterez** pas » ou « Une lumière un peu plus tamisée **aurait été parfaite** ».

**L'intention** : lorsque les internautes communiquent leurs souhaits de revenir ou pas dans un restaurant, nous parlons plutôt d'intentions. Benamara *et al.* (2017) utilisent le terme *intent* (en anglais) qui recouvre entre autres les désirs, les préférences et les intentions. Selon les auteurs, les intentions sont les actions qu'un humain va entreprendre pour satisfaire ses désirs. Les intentions montrent ainsi un engagement volontaire du locuteur, une action engagée mais cette fois par le locuteur lui-même ; deux indices lexicaux retenus sont les verbes au futur et le préfixe verbal "re-" utilisé généralement pour l'itération d'une action « refaire », « revenir », « renouveler », « retourner » etc.

**La description** : si les trois éléments précédents représentent l'évaluation exprimée par un locuteur, la description concerne plutôt les informations neutres associées à l'expérience vécue, qui ont peu de rapport avec une évaluation : des éléments comme la raison pour laquelle les consommateurs se rendent dans ce restaurant, les personnes qui les accompagnent, etc. Par exemple : « Nous avons dégusté en entrée une raviole au bœuf pour les uns, tartare de saumon pour les autres. », « Un choix pour un dîner en famille, afin de fêter les 18 ans de notre dernière fille. » et « J'avais réservé pour 21 heures. ».

# 3 Méthodologie

Afin de détecter automatiquement l'évaluation modélisée à travers les 6 catégories dans le corpus d'avis, nous avons suivi plusieurs étapes (voir la Figure 1). Les avis ont été d'abord prétraités, segmentés et transformés en vecteurs. Pour l'apprentissage automatique supervisée, nous avons utilisé *Scikit-learn* (Pedregosa et al., 2011). Nous avons choisi de tester trois algorithmes de classification supervisée qui sont couramment utilisés pour ces types de tâches : *Naïve Bayes*, *Support Vector Machine (SVM)* et *Logistic Regression*. Chacun de ces algorithmes produit un modèle global qui effectue une prédiction parmi les six possibilités. Notre objectif est donc d'obtenir le meilleur score de classification.

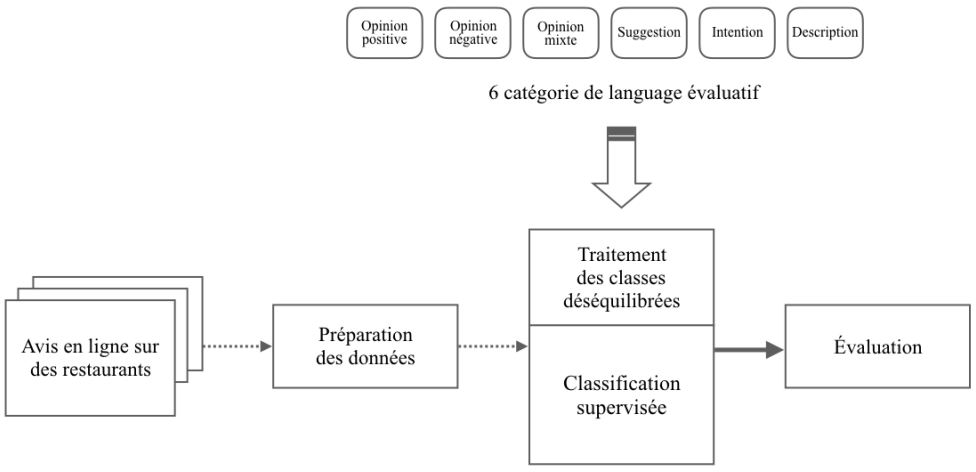



FIGURE 1: Schéma de la chaîne de traitements

## 3.1 Données

Le corpus a été collecté sur le site LaFourchette.com [2]. 200 avis par restaurant ont été extraits (au total 21 158 avis sur 126 restaurants). Le corpus a été segmenté en phrases selon les signes de ponctuation et annoté manuellement par trois annotateurs. Une des 6 catégories a été attribuée à chaque phrase segmentée. Le corpus ainsi annoté contient 2 943 phrases avec une moyenne de 10 mots par phrase. Si une phrase comprend deux catégories, notamment une opinion et une autre catégorie (intention ou suggestion), la phrase est annotée selon l'autre catégorie. Nous justifions ce choix par le fait que, mis à part les opinions, toutes les autres catégories sont peu représentées dans notre corpus. Ce problème de classes déséquilibrées est détaillé dans la section 4.3. L'accord inter-annotateur a été calculé en utilisant la mesure Kappa de Fleiss (Fleiss, 1981). Nous avons obtenu un score de 0,90, considéré comme 'presque parfait' selon l'échelle de Landis et Koch (1977). Si l'on observe la répartition de chaque catégorie (cf. Figure 2), les opinions positives représentent 68% de toutes les évaluations annotées, alors que les descriptions ou les intentions en font que 3% et 4%. Ainsi, la distribution des classes est fortement déséquilibrée. Dans la section 4.3, nous montrerons comment le problème de cette disproportion a été résolu.

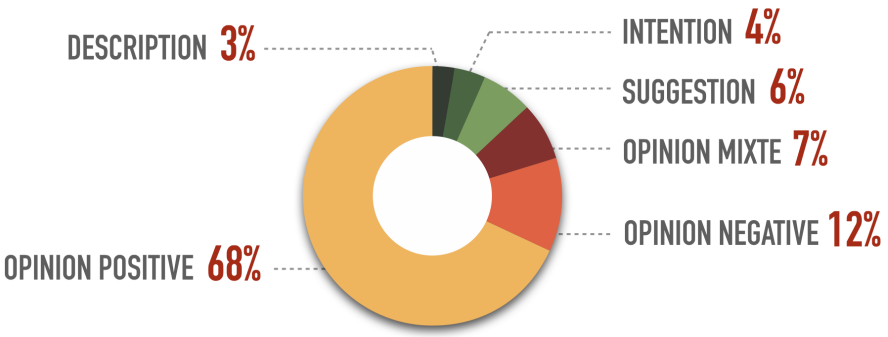



FIGURE 2: La répartition des classes dans le corpus de référence

## 3.2 Préparation des données textuelles

**Pré-traitements** : nous avons effectué le pré-traitement du corpus de la façon suivante.

- Passage des mots en minuscules ;
- Conversion des nombres par « NUM », « € » par « euros » et « % » par « pourcent » ;
- Remplacement des émoticônes par « emoPOS » ou « emoNEG » selon la polarité [3] ;
- Segmentation en phrases en suivant les signes typographiques (. ! ?) ;
- Suppression de la ponctuation. Cette étape doit être effectuée après le remplacement des émoticônes et la segmentation en phrases. Cet ordre est obligatoire pour que les émoticônes et les signes typographiques ne soient pas supprimées avant ;
- Transformation des abréviations, ce qui transforme « resto » en « restaurant » et « min » en « minutes » ;
- Lemmatisation en utilisant *TreeTagger* (Schmid 1994) ;
- Correction orthographique porte sur les mots inconnus [4] apparaissant plus que 2 fois dans le corpus.

2. La Fourchette, `https://www.lafourchette.com`
3. D'une manière générale, les émoticones sont peu présents dans notre corpus et sont toujours porteurs d'une polarité.
4. Considéres inconnus par le lemmatiseur.

**Représentation vectorielle des mots** : nous avons employé *CountVectorizer* et *TfidfVectorizer* de Scikit-learn. Nous avons testé deux paramètres dont ces méthodes disposent, *n_gram* et *max_feature*.

**Extraction de caractéristiques** :nous avons utilisé une variété de traits allant de simples traits de surface (e.g., le nombre total de caractères, de mots et la longueur moyenne de ces derniers) à des traits plus complexes comme la catégorie morphosyntaxique jugée pertinente (e.g., verbe au conditionnel, adjectif et pronom possessifs, chiffre etc.) proposée par *TreeTagger*, une valeur de la polarité et de la subjectivité identifiées par *TextBlob* [5], les mots d'opinions négatifs identifiés aussi par *TextBlob* (e.g., « déception », « désagréable », « cher », « bruyant »), la négation ainsi que la fréquence de la conjonction « mais ».

## 3.3 Traitement des classes déséquilibrées

Les travaux sur la classification de texte supposent en général une distribution équilibrée de données entre les classes (Moraes *et al.*, 2013 ; Vinodhini *et al.*, 2013 ; Zhang *et al.*, 2011 ; Ye *et al.*, 2009 ; Prabowo *et al.*, 2009 ; Tan *et al.*, 2009 ; Pang *et al.*, 2008 ; Dave *et al.*, 2003 ; Pang *et al.*, 2002). Cependant, dans la pratique, la distribution des classes est souvent asymétrique. Potts (2011) a montré que les notes des avis sur internet sont biaisées en faveur du pôle positif. Jurafsky (2014) explique ce phénomène par le principe de *Pollyanna*, une tendance générale des personnes à préférer l'information positive. La disproportion est donc un fait inévitable. À cause de ce déséquilibre, la classification est orientée en faveur de la classe majoritaire alors que les informations provenant des classes minoritaires ne sont pas prises en compte. Afin de gérer le déséquilibre entre les 6 catégories annotées, nous avons adopté différentes stratégies que l'on peut rassembler en deux groupes principaux : les stratégies d'échantillonnage et les stratégies algorithmiques (cf. le Tableau 1). Les stratégies d'échantillonnage consistent à dupliquer les observations de la classe minoritaire (*sur-échantillonnage*) ou à enlever celles de la classe dominante (*sous-échantillonnage*). Nous avons choisi le sur-échantillonnage car la classe minoritaire est sous-représentée dans les données annotées. Trois techniques de sur-échantillonnage disponibles dans *Imbalanced-learn* [6], (Lemaître *et al.* 2017) ont été utilisées : *Random Over-sampling*, *SMOTE* et *ADASYN*. Les stratégies algorithmiques consistent à pénaliser une classe sur-représentée ou à employer les méthodes ensemblistes. Il est possible de pénaliser une classe majoritaire en ajustant le paramètre *class_weight* au mode équilibré (‘*balanced*’), le poids d'une classe étant pondéré en fonction de la proportion des classes. Les méthodes ensemblistes (*Bagging*, *Boosting*) permettent également de classifier des données déséquilibrées mais elles n'ont pas été testées dans cette étude.

| **Stratégies d'échantillonnage** | **Stratégies algorithmiques** | |
|---|---|---|
| [Sur-échantillonnage] | Pénalisation | Méthodes ensemblistes |
| Random Sampling<br>SMOTE<br>ADASYN | Paramètre<br>*class_weight* | Bagging<br>Boosting<br>Simple vote |

TABLE 1: Différentes stratégies pour traiter des classes déséquilibrés

5. Une librairie Python pour traiter les données textuelles, `https://textblob.readthedocs.io/en/dev/index.html`

6. Un package Python pour traiter les données disparates (Lemaître *et al.* 2017).

## 3.4 Expérience et résultats

Pour apprendre les six étiquettes proposées, trois algorithmes de classification supervisée ont été utilisés : Naïve Bayes, Support Vector Machine (SVM) et Logistic Regression. Les expériences ont été effectuées avec une procédure de grille de recherche (*GridSearch* [7]) en utilisant une validation croisée à 5 plis. Lors des expériences, nous avons testé 10% de données. Nous y avons également associé les stratégies d'échantillonnage et d'algorithmiques (section 4.3) sur les données d'apprentissage, à l'exception de Naïve Bayes, qui ne possède pas de paramètre *class_weight*. 11 combinaisons différentes ont été testées. L'évaluation a été faite en termes de précision, rappel et F-mesure (macro moyenne). La macro F-mesure, donnant un poids identique à chaque catégorie sans tenir compte de leur taille, est donc pertinente lorsqu'il s'agit d'évaluer des données déséquilibrées (Müller & Guido, 2016). Le sur-échantillonnage de ADASYN utilisé avec l'algorithme SVM semble donner la meilleure macro F-mesure (0,79). Les résultats sont présentés dans le Tableau 2. La performance de Naïve Bayes est plus faible que les deux autres algorithmes, particulièrement sur les opinions mixtes. Ce résultat est dû au classifieur qui analyse les traits indépendamment les uns des autres. Cependant, les opinions mixtes étant composées d'opinions positives et négatives, il est difficile d'avoir des traits complètement indépendants. Nous remarquons également que la performance de la méthode classique de l'apprentissage automatique, qui ne prend pas en compte les classes disparates, est différente de celles des stratégies permettant d'avoir des classes équilibrées. Ces stratégies donnent un résultat fiable contrairement à la méthode classique qui est biaisée par la classe majoritaire. De plus, les caractéristiques de la classe minoritaire sont traitées comme du bruit et sont souvent ignorées.

| | **Macro F-mesure** | | | | |
|---|---|---|---|---|---|
| | **Classique** | **Random** | **ADASYN** | **SMOTE** | **Balanced** |
| **Naïve Bayes** | 0,66 | 0,59 | 0,61 | 0,60 | - |
| **SVM** | 0,76 | 0,77 | 0,79 | 0,78 | 0,76 |
| **Logistic Regression** | 0,78 | 0,72 | 0,72 | 0,72 | 0,77 |

TABLE 2: Macro F-mesure

Nous avons comparé nos résultats à ceux obtenus lors de l'atelier SemEval 2016 (Pontiki *et al.*, 2016). L'atelier a proposé une tâche de classification des polarités de phrases issues d'avis clients de restaurants (en anglais et en français), ce qui se rapproche de nos expériences. La tâche 5 (sous-tâche 1-3) portait sur l'attribution de la polarité (POSITIVE, NEGATIVE, NEUTRE) à une phrase donnée. La compétition utilisait cependant la métrique « Accuracy » (l'exactitude), désignant la proportion des données qui ont été classées correctement. Le meilleur accuracy (78,826) a été réalisé par Brun *et al.* (2016). La même tâche dans nos expériences a reçu l'accuracy de 87,797 en utilisant le sur-échantillonnage ADASYN associé à l'algorithme SVM. Ainsi nos résultats semblent robustes.

La précision et le rappel du sur-échantillonnage d'ADASYN utilisé avec l'algorithme SVM sont présentés dans le Tableau 3. Nous pouvons remarquer que les scores pour l'opinion positive sont élevés puisqu'ils se situent entre 0,88 et 0,96. A contrario, les autres catégories ont des résultats assez variés avec notamment une précision supérieure au rappel. Les mauvais résultats pour l'opinion mixte peuvent s'expliquer par l'implication des opinions positives et négatives dans le calcul. Autrement dit, une phrase d'opinion n'est pas toujours positive ou négative.

7. Il nous permet de trouver la meilleure combinaison d'hyper-paramètre d'un classifieur.

| | Positive | Négative | Mixte | Suggestion | Intention | Description |
|---|---|---|---|---|---|---|
| **Précision** | 0,90 | 0,86 | 0,47 | 0,94 | 1,00 | 1,00 |
| **Rappel** | 0,96 | 0,73 | 0,53 | 0,83 | 0,80 | 0,64 |

TABLE 3: Précision et Rappel

# 4 Conclusion

Les avis en ligne aboutissent aujourd'hui à une production abondante de données véhiculant l'évaluation du consommateur vis-à-vis de son expérience. L'analyse du corpus a montré qu'on ne peut pas se limiter à des notions d'opinion positive ou négative car d'autres informations et d'autres façons de les exprimer apparaissent dans le corpus. L'apprentissage automatique proposé tient compte de ces observations et des caractéristiques linguistiques de chaque catégorie retenue. Cependant, cette approche a montré un problème de déséquilibre entre les classes d'évaluation crées. Différentes approches de l'apprentissage supervisé tenant compte des spécificités du corpus d'une part et du déséquilibre de classes annotées d'autre part ont été présentées. Le sur-échantillonnage de ADASYN utilisé avec l'algorithme SVM donne la F-mesure de 0,88. En perspective, il serait intéressant de tester d'autres techniques comme des méthodes ensemblistes (*Bagging*, *Boosting*) qui pourraient donner des résultats satisfaisants. Ces méthodes combinent différents algorithmes afin d'optimiser les performances du classifieur global. Dans les challenges Netflix (2009), KDD-Cup (2009-2011) et Kaggle [8], ces méthodes ont démontré les meilleures performances (Zhou, 2012). Il serait intéressant également dans des travaux futurs de pouvoir mesurer la généricité de cette approche en appliquant sur d'autres types d'avis (messages plus longs, formes différentes, issus de corpus oraux etc.).

8. http://blog.kaggle.com/?s=1st+Place+Winner

# Références

Ashari A., Paryudi I., & Tjoa AM. (2013). Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool. *International Journal of Advanced Computer Science and Applications* 4(11).

Benamara F., Taboada M. & Mathieu Y. (2017). Evaluative language beyond bags of words: Linguistic insights and computational applications. *Computational Linguistics* 43(1): 201-264.

Brun, C., Perez, J. & Roux, C. (2016). XRCE at SemEval-2016 Task 5 : Feedbacked Ensemble Modelling on Syntactico-Semantic Knowledge for Aspect Based Sentiment Analysis. *Proceedings of the 10th International Workshop on Semantic Evaluation* (SemEval-2016), San Diego, CA, USA.

Brun, C. & Hagege C (2013). Suggestion mining: Detecting suggestions for improvement in users' comments. *Research in Computing Science* 70.

Gopalakrishnan V. & Ramaswamy C. (2013). Performance evaluation of sentiment mining classifiers on balanced and imbalanced dataset, *International Journal of Computer Science and Business Informatics*(IJCSBI), 6 (1), 1-8.

Gopalakrishnan V. & Ramaswamy C. (2014). Sentiment Learning from Imbalanced Dataset: An Ensemble Based Method, *International Journal of Artificial Intelligence*, vol. 12, no. 2, pp. 75-87.

Gopalakrishnan V. & Ramaswamy C. (2014). Sentiment mining Using SVM-based Hybrid classification model, *Advances in Intelligent Systems and Computing*, 246, 155-162, Springer-Verlag, Berlin, Heidelberg.

Hatzivassiloglou V. & Wiebe J. (2000). Effects of Adjective Orientation and Gradability on Sentence Subjectivity. *Proceedings of the 16th International Conference on Computational Linguistics* (COLING-2000).

Hu M. & Liu B. (2004). Mining and summarizing customer reviews. *Proceeding of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004).*

Kim SM. & Hovy E. (2004). Determining the Sentiment of Opinions. *Proceedings of the 20th International Conference on Computational Linguistics* (COLING-2004).

Jurafsky D. (2014). *The Language of Food: A Linguist Reads the Menu*. Norton.

Landis JR. & Koch GG. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.

Liu B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions.* Cambridge University Press, Cambridge.

Moraes R., Valiati JOF. & Neto WPG. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications* 40(2), 621–633.

Müller AC. & Guido S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media, CA.

Negi S., Rijke M. & Buitelaar P. (2018). Open Domain Suggestion Mining: Problem Definition and Datasets. arXiv preprint arXiv:1806.02179.

Negi S., Assoja K., Mehrotra S. & Buitelaar P. (2016). A Study of Suggestions in Opinionated Texts and their automatic Detection. *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*. 170–178, Berlin, Germany.

Negi S. & Buitelaar P. (2015). Towards the Extraction of Customer-to-Customer Suggestions from Reviews. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (EMNLP-2015).. 2159–2167, Lisbon, Portugal.

Pang B., Lee L. & Vaithyanathan S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *Proceeding of 2002 conference on empirical methods in natural language*, Association for Computational Linguistics. 79–86, Philadelphia, US.

Pedregosa F., Varoquaux G., Gramfort A. *et al*. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, Journal of Machine Learning Research.

Polanyi L. & Zaenen A. (2004). Contextual valenceshifters. *Proceedings of AAAI Spring Symposiumon Exploring Attitude and Affect in Text*, 106–111.

Pontiki M., Galanis D., Papageorgiou H. *et al.* (2016). SemEval-2016 Task 5: Aspect Based Sentiment Analysis. *Proceedings of the 10th International Workshop on Semantic Evaluation* (SemEval-2016), San Diego, CA, USA.

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., & Androutsopoulos, I. (2015). SemEval -2015 Task 1 : Aspect based sentiment analysis. *Proceedings of the 9th International Workshop on Semantic Evaluation* (SemEval-2015), 486–495, Denver, CO, USA.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2014). Semeval-2014 Task 4: Aspect based sentiment analysis. *Proceedings of the 8th International Workshop on Semantic Evaluation* (SemEval-2014), Dublin, Ireland.

Potts C. (2011). "On the Negativity of Negation." *Proceedings of SALT 20:* 636–59.

Riloff E., Patwardhan S. & Wiebe J. (2006). Feature Subsumption for Opinion Analysis. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP-2006), Sydney, Australia.

Riloff E. & Wiebe J. (2003). Learning Extraction Patterns for Subjective Expressions. *Proceedings of Conference on Empirical Methods in Natural Language Processing* (EMNLP-2003), Sapporo, Japan.

Schmid H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Turney PD. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In P. Isabelle (Ed.), *Proceeding of association for computational linguistics 40th anniversary meeting*, ACL, 417–424. Philadelphia, PA, USA

Vásquez C. (2014). *The discourse of online consumer reviews*. London: Bloomsbury.

Wiebe J., Wilson T., Bruce R., Bell M. & Martin M. (2014). Learning Subjective Language. *Computational Linguistics*. 30(3): 277-308.

Wilson T., Wiebe J. & Hwa R. (2004). Just How Mad Are You? Finding Strong and Weak Opinion Clauses. *Proceedings of National Conference on Artificial Intelligence* (AAAI-2004).

Ye Q., Zhang Z. & Law R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications*, 36 (3), 6527-6535.

Yu H. & Hatzivassiloglou V. (2003). Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. *Proceeding of Conference on Empirical Methods in Natural Language Processing* (EMNLP-2003).

Zhang Z., Ye Q., Zhang Z. & Li Y. (2011). Sentiment classification of Internet restaurant reviews written in Cantonese. *Expert Systems with Applications*, 38 (6), 7674-7682.

# Outiller une langue peu dotée grâce au TALN : l'exemple du corse et de la BDLC

Laurent Kevers Florian Guéniot A. Ghjacumina Tognotti Stella Retali-Medori
UMR CNRS 6240 LISA, Università di Corsica - Pasquale Paoli
Avenue Jean Nicoli, 20250 Corte, France
kevers_l,gueniot_f,tognotti_a,medori_e@univ-corse.fr

RÉSUMÉ

Nos recherches sur la langue corse nous amènent naturellement à envisager l'utilisation d'outils pour le traitement automatique du langage. Après une brève introduction sur le corse et sur le projet qui constitue notre cadre de travail, nous proposons un état des lieux concernant l'application du TAL aux langues peu dotées, dont le corse. Nous définissons ensuite les actions qui peuvent être entreprises, ainsi que la manière dont elles peuvent s'intégrer dans le cadre de notre projet, afin de progresser vers la constitution de ressources et la construction d'outils pour le TAL corse.

ABSTRACT

**Tooling up a less-resourced language with NLP : the example of Corsican and BDLC**

Our research on the Corsican language naturally leads us to consider the use of NLP tools. After a brief introduction on Corsican and the project that constitutes our working environment, we propose an overview about the use of NLP for less-resourced languages, including Corsican. We then define the actions that can be undertaken, as well as how they can be integrated into our project, in order to progress towards the constitution of resources and the construction of tools for the Corsican NLP.

MOTS-CLÉS : langues peu dotées, corse, ressources linguistiques, lemmatisation, POS.

KEYWORDS: less-resourced languages, Corsican, linguistic resources, lemmatisation, POS.

## 1 La langue corse et sa diffusion numérique

Le corse est une langue issue du latin et s'inscrit dans l'ensemble italo-roman ; il a connu divers contacts et influences linguistiques. Il présente en particulier des affinités avec le toscan médiéval liées spécialement à la domination pisane (IX$^{ème}$-XIII$^{ème}$ s.). Le toscan constitue le superstrat du corse, et celui-ci connaît des emprunts à d'autres variétés italo-romanes voire romanes ainsi qu'aux langues germaniques et à l'arabe.

Sur le plan dialectal quatre, voire cinq aires sont identifiables (Dalbera-Stefanaggi, 2002, 2007), et l'aire méridionale extrême franchit même les frontières de la Corse puisqu'elle se prolonge en Gallura, dans le nord de la Sardaigne. Ces cinq aires dialectales constituent toutefois un *continuum* et ne font pas obstacle à intercompréhension entre les locuteurs des diverses variétés voire avec les variétés centrales et méridionales de l'Italie.

Les affinités génétiques et historiques que le corse a entretenues pendant de nombreux siècles avec le toscan en ont fait, avec l'italien qui lui a succédé, la langue d'écriture des insulaires depuis le Moyen

Âge jusqu'à l'émergence d'une écriture consciente en langue corse au XIX$^{e}$ s. L'orthographe du corse est, de ce fait et avec des adaptations, fondée sur le système graphique de l'italien [1]. Toutefois, et malgré la mise en œuvre d'une approche polynomique permettant d'englober l'ensemble des variantes dialectales (Marcellesi, 1984), l'écriture de la langue n'est pas normée ce qui implique une certaine difficulté pour son traitement automatique.

Aujourd'hui en rapport de diglossie avec le français, l'usage du corse tend à régresser et le développement d'outils est nécessaire pour sa préservation, sa valorisation, sa transmission et sa promotion [2]. Une politique au service de la langue corse est active sur le territoire insulaire notamment pour son développement par les nouvelles technologies. Si plusieurs outils et ouvrages existent pour la description linguistique des parlers corses ou encore pour leur apprentissage, l'inscription du corse dans les humanités numériques reste lacunaire [3].

En particulier, les sites et applications relatifs à la traduction, au vocabulaire et à la syntaxe contiennent peu de données en confrontation à la richesse et la complexité de la langue corse. Cette richesse se retrouve en revanche sur les bases de données telles que la *Banque de Données Langue Corse* (BDLC) et *Infcor* [4]. Cette dernière a été conçue dans un cadre associatif par l'ADECEC qui œuvre dans le domaine de la langue et la culture corses ; elle a été déclinée sous la forme d'une application smartphone mise à disposition du grand public et propose de nombreuses fiches lexicales avec des modes d'interrogation multicritères [5]. Concernant la BDLC, il s'agit d'un outil conçu dans un contexte scientifique associé à la réalisation d'un atlas linguistique, le NALC, comme nous allons l'évoquer.

## 2 Le projet *Banque de Données Langue Corse* (BDLC)

Le *Nouvel Atlas Linguistique et ethnographique de la Corse* (NALC) a été conçu par le CNRS en 1975 et confié à Marie-José Dalbera-Stefanaggi en 1981 à l'ouverture de l'Université de Corse. En 1986, répondant à une demande de l'Assemblée Régionale de Corse, a été créée la *Banque de Données Langue Corse* (BDLC) qui s'est articulée tout naturellement avec l'atlas projeté [6].

Le NALC-BDLC accueille des données linguistiques en lien avec des savoir-faire et des traditions culturelles corses sur l'ensemble du territoire insulaire. Lors d'enquêtes sur le terrain auprès de locuteurs locaux, ces données sont collectées grâce à des questionnaires thématiques [7] constitués de listes de mots en français [8] : par exemple « la vigne », « le cep », « tailler la vigne », « le tonneau ». A partir d'une question individuelle telle que *« Cumu si dici* "tailler la vigne" *in corsu ? »* (fr. :

1. Cf. notamment Retali-Medori (2015) pour une synthèse sur la question.
2. Selon les recommandations de (UNESCO, 2003).
3. La représentation de cette langue sur le web peut se retrouver sous diverses formes : interfaces en langue corse (les moteurs de recherche *Qwant* et *Google*, ou des réseaux sociaux tels que *Facebook*), sites ou applications tels que *Google Translate* ou *Wiktionnaire*, applications pour smartphone (orientées tourisme, éducation ou traduction), ressources pédagogiques en ligne par le biais de sites d'apprentissage de la langue corse, blogs relatifs à la langue et à la culture corses.
4. *Banca di dati di a lingua corsa* : `http://infcor.adecec.net`
5. Même si l'ensemble des fiches lexicales nécessiterait une révision des informations relatives à la variation, aux signifiés, à la morphologie et à l'étymologie, le matériel contenu dans cette base est incontournable et pourrait aider aux développements de nouveaux outils.
6. Une synthèse de l'histoire du projet est présentée par Dalbera-Stefanaggi & Retali-Medori (2015). Le programme est dirigé depuis 2015 par S. Retali-Medori. Une collection de semi-vulgarisation intitulée *Detti è Usi di paesi, matériaux et analyses extraits de la Banque de Données Langue Corse* a en outre été créée en 2006 autour du NALC.
7. Les thèmes développés dans la BDLC pour les recueils lexicaux sont : l'élevage, l'agriculture, l'homme, la maison et la vie quotidienne, la nature, le village ou la ville et les croyances.
8. Les questionnaires ont été créés au début du programme par le biais d'enregistrements préalables réalisés dans l'île sur différents domaines techniques ou culturels. A partir de leurs transcriptions, a été établie la liste des mots, des signifiés, nommée en d'autres termes par M. J. Dalbera-Stefanaggi le « responsaire » (Dalbera-Stefanaggi, 1992, p. 397).

« Comment dit-on “tailler la vigne” en corse ? »), la collecte des traductions correspondantes en corse est permise et un entretien semi-dirigé entièrement en corse et relatif aux pratiques s'engage afin de recueillir aussi des ethnotextes (témoignages). Ces données sont ensuite traitées et analysées –sur le plan linguistique– et mises en ligne sur le site `http://bdlc.univ-corse.fr`.

Si cette base de données constitue un réel outil pour le développement du TAL appliqué au corse, une des difficultés majeures provient de la riche variation caractérisant la langue corse. Selon les signifiés, des variations lexicales importantes sont attestées : par exemple pour désigner l'acte d'épamprer la vigne, 25 lemmes ont été collectés. Ces lemmes vont à leur tour connaître des transcriptions variables notamment en conséquence de la non-normalisation de la langue corse et d'une production réalisée par différents transcripteurs au cours des 30 années d'existence du programme. Les différents choix d'écritures répondent à des objectifs tels que :

- valoriser la variation : par exemple pour nommer « la jarre », selon la prononciation dans les localités enquêtées, nous trouverons les formes *cerra* et *gerra* issues du même étymon ;
- indiquer graphiquement dans les textes l'aperture des voyelles des proparoxytons (*tróvula* : « écuelle » ; *còmpulu* : « abri » ; *pèrgula* : « treille » ; *tépidu* : « tiède ») ou l'accentuation des hiatus (*durmìa* : « il dormait ») ;
- représenter des enclitiques tels que *fanne* (« en faire ») correspondant à *fà* (« faire ») + *ne* (« en »).

# 3 Réingénierie et modernisation des outils de la BDLC

En 1986, une première version de l'application BDLC a été développée en *standalone* en utilisant le système de gestion de bases de données 4D. Cette application proposait une fonction de recherche basée sur un ensemble fermé de mots en français, organisés par thèmes. Les réponses obtenues correspondaient à une ou plusieurs traductions en corse et comportaient la forme phonique, la forme graphique, le lemme et la localité d'origine de la forme, et pouvaient être illustrées par des sons ou des photos. Des données morphologiques (flexion et découpage morphématique) et étymologiques étaient également disponibles. La géolocalisation des réponses permettait la visualisation des variations dialectales à l'échelle de l'île avec des cartes de lemmes ou des variantes phoniques.

Une boîte à outils proposait plusieurs modules complémentaires : un module morphologique donnait accès à des tableaux verbaux et nominaux ainsi qu'à la segmentation des mots en morphèmes ; un module étymologique donnait la liste des étymons des termes corses, ou la liste des continuateurs corses d'un étymon, ainsi qu'un fichier morphématique étymologique (découpage des étymons en morphèmes). Il était aussi possible d'établir des requêtes de phonétique diachronique ou de morphologie diachronique (segmentation ou reconstruction en morphèmes des étymons).

L'obsolescence de l'application, et le fait qu'elle soit *standalone*, impliquait la coexistence de différentes versions selon les postes de travail. Une refonte complète a donc été réalisée en tirant parti des technologies web actuelles : PHP et Javascript pour l'applicatif, MySQL pour la base de données. Cette nouvelle version, reprenant en substance une grande partie des fonctionnalités de son aïeule, a connu différentes versions successives. En 2016, la dernière version a été développée en interne par Florian Guéniot (IGE, CNRS) et Aloïs Beck (Alternant, Université de Corse) dont l'objectif a été double : recoder le moteur de l'application afin de le rendre évolutif et modulaire, et moderniser l'interface graphique afin de le rendre adaptable aux supports de plus en plus variés. Cette refonte complète, y compris de la structure de la base de données, a été l'occasion de faciliter l'accès aux données pour les opérateurs par l'ajout d'un module de recherche multicritères. La base de données comporte actuellement 4.888 questions, 108.867 réponses et 1.288 ethnotextes.

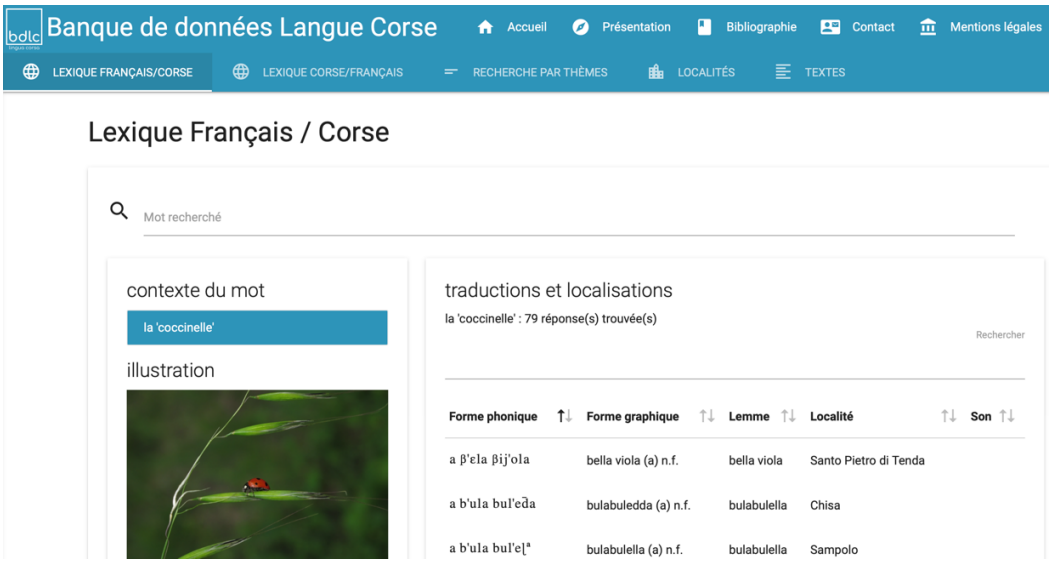



FIGURE 1 – Interface de consultation de la BDLC

# 4 Vers un TAL corse

Suite à ces évolutions techniques qui modernisent et pérennisent le projet BDLC, grâce à son corpus à la fois authentique et reflet de la complexité et de la richesse de la langue, et étant donné son enrichissement constant, la *Banque de Données Langue Corse* peut constituer un terreau intéressant afin d'œuvrer à la construction de ressources et d'outils pour le traitement automatique du corse.

## 4.1 État de la question

À notre connaissance, il n'existe que très peu de ressources et d'outils pour le TAL corse. Le rapport de l'ELDA de 2014 sur les ressources linguistiques consacrées aux langues de France (Leixa *et al.*, 2014) recense 93 ressources pour le corse. Plus d'un tiers de celles-ci sont des enregistrements issus du projet BDLC. Les deux tiers restants sont constitués de documents divers : blogs, articles scientifiques, sites institutionnels, sites de journaux, etc. On y retrouve également quelques lexiques, dont *Infcor* ou le *Wiktionnaire corse*, déjà mentionnés précédemment. En addition à cet inventaire, on peut touver quelques autres contributions, au rang desquelles figure le réseau sémantique *BabelNet* (Navigli & Ponzetto, 2012) qui propose un certain nombre d'éléments en corse, ou encore les ressources corses constituées au sein du *Crúbadán Project* (Scannell, 2007). À l'exception de ces dernières, la majorité des ressources disponibles ne sont pas directement exploitables en TAL.

Le corse rentre dans la catégorie des langues dites « peu ou mal dotées », ou encore « minoritaires ». Ces langues constituent un domaine de recherche actif. La conférence TALN a accueilli plusieurs événements dédiés à cette question, entre autres le workshop « Traitement automatique des langues minoritaires et des petites langues » (Streiter, 2003), ainsi que les workshop TALaRE, « Traitement Automatique des Langues Régionales de France et d'Europe » (Morin & Estève, 2013; Vergez-Couret *et al.*, 2015). De même, la conférence LREC a proposé de multiples workshops, dont les plus récents sont SaLTMiL (Alegria *et al.*, 2010; De Pauw *et al.*, 2012) et CCURL (Pretorius *et al.*, 2014; Soria *et al.*, 2016, 2018). Dernièrement, la revue TAL a également sorti un numéro thématique sur le sujet (Bernhard & Soria, 2018). La place de la langue corse dans ces publications est cependant faible.

## 4.2 Objectifs et moyens

Lorsque l'on désire initier ou améliorer le traitement informatique d'une langue peu ou mal dotée, il est logique de créer les ressources de base avant de s'attaquer aux outils. Ces ressources sont habituellement constituées de lexiques et/ou de corpus, annotés ou non, monolingues ou parallèles. Les outils sont souvent élaborés suivant une complexité croissante. On partira par exemple d'un détecteur de langues, déjà utile lors de la phase de constitution des corpus, pour développer ensuite des composants d'analyse morphosyntaxique et lexicale, pour enfin aller vers des applications de plus haut niveau telles que la correction orthographique ou la traduction automatique. Un aperçu d'actions à entreprendre pour améliorer la capacité digitale des langues est proposé par Ceberio Berger *et al.* (2018). Notre approche rejoint leurs recommandations.

Dans le cadre de notre projet, nous avons décidé d'avancer sur différents points en parallèle. En termes d'objectifs, nous désirons disposer en premier lieu :

- d'un dictionnaire électronique exploitable pour le TAL ;
- d'une interface de consultation de textes capable de générer des concordances et de répondre à des requêtes incluant des critères linguistiques (interrogation, éventuellement combinée, sur les formes, les lemmes, les catégories grammaticales et flexionnelles, etc.) ;
- d'un outil de détection de langue ;
- d'un outil d'annotation morphosyntaxique.

Du point de vue technique, ces objectifs impliquent :

- la constitution initiale et l'enrichissement progressif du dictionnaire électronique ;
- la définition d'une procédure de lemmatisation et son application à une base textuelle à intégrer dans l'interface de consultation (en l'occurrence, les ethnotextes issus de la BDLC) ;
- la mise en place de cette interface ;
- la création de corpus corses à des fins d'entraînement (e.a. pour la détection de langue et l'annotation morphosyntaxique) ;
- la mise au point des outils de détection de langue et d'annotation morphosyntaxique.

Nous avons donc en premier lieu construit une version initiale du dictionnaire à partir d'un export de la BDLC, ainsi que de quelques ajouts extérieurs en ce qui concerne les verbes [9]. Les données ont été organisées selon le format des dictionnaires du LADL [10] (Gross, 1989; Courtois, 1990; Silberztein, 1993). Cette première ressource permet l'initialisation du processus de lemmatisation. Elle sera, en retour, enrichie à l'issue de celui-ci. Actuellement, le dictionnaire compte 20.875 formes, dont 17.860 formes simples (se rapportant à 10.224 lemmes) et 3.015 formes composées (se rapportant à 2.244 lemmes). Lorsque ce dictionnaire est appliqué à notre corpus d'ethnotextes représentant environ 160.000 formes, dont un peu moins de 15.000 uniques, environ 49 % des occurrences sont reconnues. Pour ces éléments, plusieurs analyses concurrentes peuvent coexister (ambiguïté lexicale) et l'analyse correcte peut éventuellement être absente (incomplétude du dictionnaire). Notons encore qu'un traitement des formes non reconnues les plus fréquentes permet d'améliorer rapidement la couverture : les 20 premièrs de ces éléments couvrent pas moins de 31 % du total des formes inconnues. À terme, ce dictionnaire constituera une ressource directement exploitable en TAL.

Le deuxième chantier entammé est celui de la lemmatisation. Cette tâche répond à un triple objectif. D'une part, permettre une interrogation des textes sur des critères linguistiques, ainsi qu'une restitution des résultats sous la forme de concordances. D'autre part, nous visons la constitution d'un

9. Les principales formes de *esse* (« être »), *avè* (« avoir »), *andà* (« aller »), *dà* (« dire »), *fà* (« faire »), *stà* (« être », état).

10. Les entrées sont enregistrées dans des fichiers textes sous le format suivant : `forme,lemme.codes_grammaticaux_sémantiques:code_flexionnels/commentaire`.

corpus annoté permettant, dans le futur, de réaliser des apprentissages artificiels (e.a. un étiqueteur morphosyntaxique, cf. *infra*). Enfin, comme déjà exposé, la lemmatisation permettra l'élaboration progressive d'un lexique électronique exploitable en TAL. La définition de la procédure de lemmatisation s'appuie sur l'expérience du projet GREgORI [11]. Celui-ci a, depuis des années, mis au point des méthodologies et des outils pour l'aide à la lemmatisation de textes en grec ancien et dans les principales langues de l'orient chrétien (Kevers & Kindt, 2004; Kindt, 2018). Ces outils, qui peuvent être transposés pour le traitement du corse, permettent une automatisation partielle de la lemmatisation (Kindt, 2012). En pratique, nous avons défini le déroulement de la lemmatisation en deux grandes étapes (figure 2) : l'« étude lexicographique », durant laquelle les formes inconnues du dictionnaire sont ajoutées à celui-ci, et la « désambiguïsation », qui permet de ne conserver qu'une et une seule analyse pour chaque forme du texte. Lors de cette seconde étape, des ajouts au dictionnaire sont encore envisageables. C'est le cas des formes qui n'ont pas été prises en compte lors de l'étude lexicographique, car déjà présentes dans le dictionnaire, mais pour lesquelles l'analyse adéquate n'est pas encore proposée. Concrètement, ces traitements sont mis en œuvre au moyen du logiciel Unitex [12] (Paumier, 2016). Le processus de lemmatisation représente un effort conséquent, mais il permet de travailler au plus près des données et d'en améliorer la qualité, ce qui est important en vue de leur consultation à des fins scientifiques et de leur utilisation pour l'apprentissage artificiel.

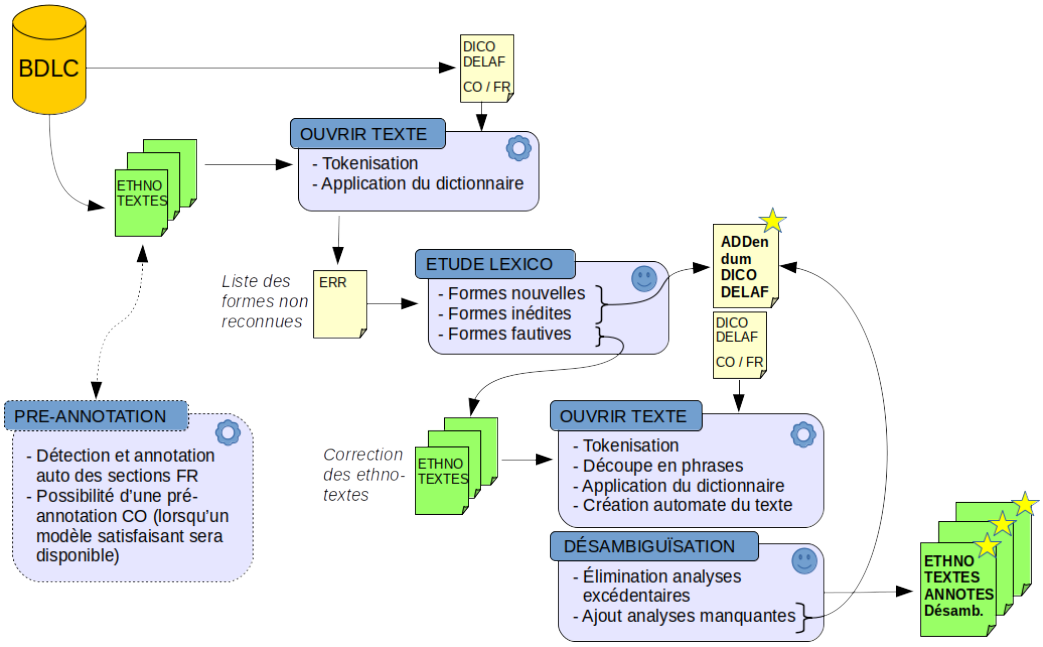



FIGURE 2 – Processus de lemmatisation

Cette procédure est en cours de précision sur différents points importants liés aux variations dialectales et à la non normalisation : choix du lemme entre les différentes versions attestées selon les régions, gestion des variations dûes à l'utilisation variable des accents. Différentes actions sont également entreprises afin de rassembler des données lexicales supplémentaires, extérieures à la BDLC, susceptibles de venir enrichir la version initiale du dictionnaire. En effet, suite à nos premières analyses, nous avons constaté qu'il serait intéressant, en termes d'efficacité, de démarrer la lemmatisation avec

11. Centre d'études orientales, Institut Orientaliste de l'université de Louvain-la-Neuve (Belgique) : `https://uclouvain.be/fr/instituts-recherche/incal/ciol/gregori-project.html` ; assisté par le Centre de traitement automatique du langage : `https://uclouvain.be/fr/instituts-recherche/ilc/cental`.

12. `https://unitexgramlab.org`

un lexique électronique plus couvrant que celui dont nous disposons actuellement.

Au delà du traitement des ethnotextes de la BDLC, nous travaillons également à la constitution de plus gros volumes de textes dans le but de créer des corpus, mono- ou multilingues (3ème chantier). La récolte de documents sur le web et la recherche de corpus déjà constitués font partie de nos préoccupations actuelles. Le recours à des techniques de *crowdsourcing* appliquées à la construction de ressources linguistiques (Millour & Fort, 2018) pourra également être envisagé, non seulement pour produire des corpus annotés en parties du discours, mais aussi des lexiques et corpus multilingues. Ces ressources pourront nous être utiles pour la mise au point d'outils (annotation morphosyntaxique, détection de langue, cf. *infra*), mais aussi à plus long terme, nous permettre de travailler à l'élaboration d'applications de haut niveau telles que la traduction automatique [13], principalement vers l'italien et le français.

L'entraînement et l'utilisation d'un étiqueteur morphosyntaxique, tel que le Tree Tagger (Schmid, 1994), le Stanford Part-Of-Speech Tagger (Toutanova *et al.*, 2003) ou encore d'un outil tel que Wapiti (Lavergne *et al.*, 2010), est provisoirement laissé de côté. Au fur et à mesure de la production de textes lemmatisés, des modèles de plus en plus complets pourront cependant être entraînés. Dès qu'un niveau de qualité suffisant sera atteint, l'étiqueteur pourra venir effectuer une pré-annotation en support du processus de lemmatisation (figure 2). Parallèlement, nous testons également la possibilité d'exploiter un étiqueteur déjà entraîné pour une langue proche et mieux dotée, en l'occurrence l'italien à l'aide du Tree Tagger. Ce type de démarche a entre autres déjà été expérimentée avec des résultats intéressants pour l'alsacien et l'occitan (Bernhard & Ligozat, 2013; Vergez-Couret, 2013; Bernhard *et al.*, 2018).

Enfin, la quatrième action, en cours actuellement, concerne la mise au point d'un détecteur de langue capable de reconnaître le corse. En plus de l'intérêt intrinsèque de cet outil, il nous sera à nouveau utile pour les tâches déjà initiées, e.a. la lemmatisation (figure 2) et la constitution de corpus. Nous nous intéressons dans un premier temps à la détection de la langue principale du texte, mais visons également la détection de segments de différentes langues à l'intérieur du document (les textes de la BDLC mixent parfois le corse et le français). Pour ce travail, nous nous appuyons sur l'étude de Jauhiainen *et al.* (2018) et avons commencé à tester et apprendre des modèles pour différents outils, sur la base des premiers corpus récoltés. Au vu des premiers résultats, nous devrions être en mesure de nous rapprocher assez rapidement des performances au niveau de l'état de l'art.

# 5 Conclusion

Nos recherches sur la langue corse nous amènent naturellement à envisager l'utilisation d'outils pour le traitement automatique du langage. L'état des lieux de ce domaine, et plus particulièrement de son application au corse, nous a révélé le manque de ressources et d'outils en la matière. Nous pensons cependant que notre projet peut constituer un terreau intéressant afin d'œuvrer à leur construction. Nous avons donc esquissé les grandes lignes des actions à entreprendre pour nous faire progresser vers la mise en œuvre du TAL corse : constitution de corpus annotés par la lemmatisation semi-automatique, création concomitante d'un lexique exploitable pour le TAL, expérimentation d'outils de base, dont un détecteur de langues et un analyseur morphosyntaxique. Les premières étapes ont été entamées, les plus ambitieuses suivront progressivement. À plus long terme, une application telle que la traduction automatique est envisagée. Les travaux réalisés sur d'autres langues régionales ou peu dotées, dont nous avons dressé un bref aperçu, nous guideront dans ce cheminement.

13. Un premier essai, non satisfaisant, utilisant le *deep learning* (Tensor Flow) pour la traduction automatique, nous a montré la nécessité de disposer de données conséquentes pour cette tâche.

# Références

I. ALEGRIA, N. BEL, L. BORIN, H. LOFTSSON, F. SANCHEZ-MARTINEZ, K. SCANNELL, T. TROSTERUD, P. LANGGA, P. MEURER, S. MOSHAGEN, E. NAVAS & D. TOMAS, Eds. (2010). *7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages (LREC2010 Workshop)*.

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

BERNHARD D., LIGOZAT A.-L., MARTIN F., BRAS M., MAGISTRY P., VERGEZ-COURET M., STEIBLÉ L., ERHART P., HATHOUT N., HUCK D., REY C., REYNES 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 *Proceedings of the 11th edition of the Language Resources and Evaluation Conference (LREC'18)*, Miyazaki, Japon.

BERNHARD D. & SORIA C. (2018). Traitement automatique des langues peu dotées. *Traitement Automatique des Langues*, **59**(3).

CEBERIO BERGER K., GURRUTXAGA HERNAIZ A., BARONI P., HICKS D., KRUSE E., QUOCHI V., RUSSO I., SALONEN T., SARHIMAA A. & SORIA C. (2018). *Digital Language Survival Kit. The DLDP Recommendations to Improve Digital Vitality*. The Digital Language Diversity Project. Accessible à l'adresse <http ://www.dldp.eu/sites/default/files/documents/DLDP_Digital-Language-Survival-Kit.pdf>.

COURTOIS B. (1990). Un système de dictionnaires électroniques pour les mots simples du français. *Langue Française*, **87**, 11–22.

DALBERA-STEFANAGGI M.-J. (1992). Le Nouvel Atlas Linguistique de la Corse et son articulation sur une base de données. In *Atlanti Linguistici italiani e romanzi : esperienze a confronto. Atti del Congresso Internazionale (Palermo 3-7 Ottobre 1990)*, p. 395–402, Palermo : Centro di Studi Filologici e Linguistici Siciliani.

DALBERA-STEFANAGGI M.-J. (2002). *La langue corse*. Number 3641 in Que sais-je ? Paris : PUF.

DALBERA-STEFANAGGI M.-J. (2007). *Nouvel atlas linguistique et ethnographique de la Corse : Volume 1, Aréologie phonétique, édition revue et corrigée*. Ajaccio : Paris : Comité des travaux historiques et scientifiques - CTHS, Alain Piazzola edition.

DALBERA-STEFANAGGI M.-J. & RETALI-MEDORI S. (2015). Trente ans de dialectologie corse : autour du programme Nouvel Atlas Linguistique et Ethnographique de la Corse et Banque de Données Langue Corse. In S. RETALI-MEDORI, Ed., *Actes du colloque Tribune des chercheurs, études en linguistique*, volume 6 of *Corse d'hier et de demain - Nouvelle série*, p. 17–25, Bastia, France : Société des Sciences Historiques et Naturelles de la Corse.

G. DE PAUW, G.-M. DE SCHRYVER, M. L. FORCADA, K. SARASOLA, F. M. TYERS & P. W. WAGACHA, Eds. (2012). *8th SaLTMIL & AfLaT 2012 Workshop, Language Technology for Normalisation of Less-Resourced Languages (LREC 2012 Workshop)*.

GROSS M. (1989). La construction de dictionnaires électroniques. *Annales de Télécommunications*, **44**, 4–19.

JAUHIAINEN T., LUI M., ZAMPIERI M., BALDWIN T. & LINDÉN K. (2018). Automatic Language Identification in Texts : A Survey. *arXiv :1804.08186 [cs]*.

KEVERS L. & KINDT B. (2004). Vers un concordanceur-lemmatiseur en ligne du grec ancien. *L'Antiquité Classique*, **73**, 203–213.

KINDT B. (2012). *Traitement automatique de l'ambiguïté en grec ancien. Outils informatiques et ressources linguistiques.* Thèse de doctorat., Université catholique de Louvain.

KINDT B. (2018). Processing Tools for Greek and Other Languages of the Christian Middle East. *Journal of Data Mining and Digital Humanities*, **jdmdh :4184**. Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages,.

LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical Very Large Scale CRFs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, p. 504–513, Uppsala, Sweden : Association for Computational Linguistics.

LEIXA J., MAPELLI V. & CHOUKRI K. (2014). *Inventaire des ressources linguistiques des langues de France*. ELDA. Accessible à l'adresse <http ://www.elda.org/media/filer_public/2014/12/17/rapport_dglflf_05112014-1.pdf >.

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

MILLOUR A. & FORT K. (2018). À l'écoute des locuteurs : production participative de ressources langagières pour des langues non standardisées. *Traitement Automatique des Langues, numéro spécial sur le traitement automatique des langues peu dotées*, **59**(3).

E. MORIN & Y. ESTÈVE, Eds. (2013). *TALaRE 2013 - Traitement Automatique des Langues Régionales de France et d'Europe. Atelier TALN 2013*, Les Sables d'Olonne, France.

NAVIGLI R. & PONZETTO S. P. (2012). BabelNet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, **193**, 217–250.

PAUMIER S. (2016). *Unitex 3.1 User Manual*. Université Paris-Est Marne-la-Vallée. Accessible à l'adresse http ://releases.unitexgramlab.org/3.1/man/Unitex-GramLab-3.1-usermanual-en.pdf.

L. PRETORIUS, C. SORIA & P. BARONI, Eds. (2014). *CCURL 2014 : Collaboration and Computing for Under - Resourced Languages in the Linked Open Data Era, LREC 2014 Workshop*.

RETALI-MEDORI S. (2015). La documentation corse. In E. R. MARIA ILIESCU, Ed., *Anthologies, textes, corpus et sources des langues romanes*, number 7 in Manuals of Romance Linguistics, p. 558–564. Tübingen : De Gruyter.

SCANNELL K. P. (2007). The Crúbadán Project : Corpus building for under-resourced languages. In C. FAIRON, H. NAETS, A. KILGARRIFF & G.-M. DE SCHRYVER, Eds., *Proceedings of the 3rd Web as Corpus Workshop*, volume 4 of *Cahiers du Cental*, Louvain-la-Neuve, Belgium.

SCHMID H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

SILBERZTEIN M. D. (1993). *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*. Paris : Masson.

C. SORIA, L. BESACIER & L. PRETORIUS, Eds. (2018). *CCURL 2018 : 3rd Workshop on Collaboration and Computing for Under-Resourced Languages, Sustaining knowledge diversity in the digital age (LREC 2018 Workshop)*.

C. SORIA, L. PRETORIUS, T. DECLERCK, J. MARIANI, K. SCANNELL & E. WANDL-VOGT, Eds. (2016). *CCURL 2016 : Collaboration and Computing for Under-Resourced Languages : Towards an Alliance for Digital Language Diversity (LREC 2016 Workshop)*.

O. STREITER, Ed. (2003). *Traitement automatique des langues minoritaires et des petites langues, Actes du Workshop TALN 2003*, Batz-sur-Mer. ATALA.

TOUTANOVA K., KLEIN D., MANNING C. D. & SINGER Y. (2003). Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL 2003*, p. 173–180 : Association for Computational Linguistics.

UNESCO, GROUPE D'EXPERTS SPÉCIAL DE L'UNESCO SUR LES LANGUES EN DANGER (2003). *Vitalité et disparition des langues - UNESCO Bibliothèque Numérique*. Paris : Organisation des Nations Unies pour l'Education, la Science et la Culture. Version française accessible à l'adresse <https ://ich.unesco.org/doc/src/00120-FR.pdf>.

VERGEZ-COURET M. (2013). Tagging Occitan using French and Castillan Tree Tagger. In *Less Resourced Languages, new technologies, new challenges and opportunities*, p.6̃, Poznan, Poland.

M. VERGEZ-COURET, D. BERNHARD, A.-L. LIGOZAT, J.-M. ELOY & C. REY, Eds. (2015). *TALaRE 2015 - Traitement Automatique des Langues Régionales de France et d'Europe. Atelier de TALN 2015*, Caen, France.

# Plongements lexicaux spécifiques à la langue arabe: application à l'analyse d'opinions

Amira Barhoumi[1, 2] Nathalie Camelin[1] Chafik Aloulou[2] Yannick Estève[1]
Lamia Hadrich Belguith[2]
(1) LIUM, adresse, 72000 Le Mans, France
(2) MIRACL, adresse, CP Sfax, Tunisie
amira.barhoumi.etu@univ-lemans.fr, nathalie.camelin@univ-lemans.fr,
chafik.aloulou@fsegs.rnu.tn, yannick.esteve@univ-lemans.fr,
l.belguith@fsegs.rnu.tn

RÉSUMÉ

Nous nous intéressons, dans cet article, à la tâche d'analyse d'opinions en arabe. Nous étudions la spécificité de la langue arabe pour la détection de polarité. Nous nous focalisons ici sur les caractéristiques d'agglutination et de richesse morphologique de cette langue. Nous avons particulièrement étudié différentes représentations d'unité lexicale : token, lemme et light stemme. Nous avons construit et testé des espaces continus de ces différentes représentations lexicales. Nous avons mesuré l'apport de tels types de representations vectorielles dans notre cadre spécifique. Les performances du réseau CNN montrent un gain significatif de 2% par rapport à l'état de l'art.

ABSTRACT

**Arabic-specific embedddings : application in Sentiment Analysis**

In this article, we are interested in Arabic sentiment analysis task. We study the specificity of the Arabic language for the detection of polarity. We focus on the agglutination and morphological richness of this language. We particularly studied lexical units of different granularities : tokens, lemmas and light stems. We have built and tested continuous spaces of these lexical units. We have measured the contribution of such types of embeddings in our specific framework. The performance of the CNN network has a significant gain of 2%.

MOTS-CLÉS : Analyse d'opinion, représentation vectorielle continue, apprentissage profond, langue arabe.

KEYWORDS: Sentiment analysis, embeddings, deep learning, arabic language.

## 1 Introduction

Avec Internet et l'explosion des réseaux sociaux, un grand nombre d'internautes expriment leurs points de vue et leurs sentiments sur des entités, des produits, des personnes, *etc.* Dans ce cadre, le domaine d'analyse automatique d'opinions est en plein essor. Il consiste souvent à identifier la subjectivité et la polarité (positive, négative, neutre) d'un énoncé donné (Pang *et al.*, 2008; Yang *et al.*, 2017). Dans ce travail, nous nous intéressons à l'analyse d'opinions à partir de textes rédigés en langue arabe. Les travaux effectués dans ce domaine permettent de distinguer trois approches. La première est symbolique, elle utilise des lexiques (Abdulla *et al.*, 2014a) et des règles linguistiques

(Almas & Ahmad, 2007; Farra *et al.*, 2010). La deuxième consiste en une approche statistique qui s'appuie sur des méthodes d'apprentissage automatique (Abdulla *et al.*, 2014b; Bayoudhi *et al.*, 2015). Quant à la troisième, elle est hybride, elle utilise à la fois des lexiques et des algorithmes d'apprentissage automatique (El-Halees, 2011; Ibrahim *et al.*, 2015; Refaee & Rieser, 2016).
La recherche en analyse d'opinions a tiré profit des avancées scientifiques dans les techniques d'apprentissage profond, et plusieurs travaux ont été récemment réalisés avec ce type d'apprentissage. (Al Sallab *et al.*, 2015) teste différents réseaux profonds. (Dahou *et al.*, 2016; Barhoumi *et al.*, 2018), quant à eux, utilisent une architecture à base de réseau convolutif CNN. La majorité des réseaux neuronaux prennent comme entrée des représentations vectorielles continues (*embeddings*) de mots. L'espace de projection est un espace continu supposé préserver les similarités sémantiques et syntaxiques des mots. Les word embeddings se sont révélés être un atout fondamendal pour plusieurs tâches de traitement du langage naturel, y compris l'analyse d'opinions.

Word2vec (Mikolov *et al.*, 2013) et Glove (Pennington *et al.*, 2014) sont parmi les algorithmes de construction d'embeddings les plus répandus. Plus récemment, les embeddings contextuels *Elmo* (Peters *et al.*, 2018) et BERT (Grave *et al.*, 2018) sont apparus pour gérer à la fois les contextes linguistiques et la syntaxe/sémantique des mots. Pour la langue arabe, quelques ressources d'embeddings sont disponibles. (Dahou *et al.*, 2016) a entrainé le modèle word2vec (Mikolov *et al.*, 2013) de type CBOW sur des pages web. (Soliman *et al.*, 2017) regroupe six modèles d'*embeddings* (CBOW et Skip-gram) entrainés sur trois types de corpus différents : twitter, wikipédia et des pages web. (Barhoumi *et al.*, 2018) utilise ces embeddings de mots pré-entrainés et obtient de bonnes performances pour l'analyse d'opinions en arabe.
Néanmoins, des améliorations peuvent encore être obtenues si l'on prend en compte les nombreuses spécificités de la langue arabe. Notamment, si l'on considère que la définition d'un mot, au sens graphique, est une séquence de caractères délimitée par deux séparateurs (blanc ou autre marqueur de séparation, tel que la ponctuation), alors un mot en arabe peut avoir une structure très complexe. En effet, ce mot peut être décomposable en proclitique(s), forme fléchie et enclitique(s). Par exemple, le mot أسيعجبه /AsyEjbh/ (est-ce qu'il va lui plaire), se compose d'une particule d'interrogation ا et de futur س, de la forme fléchie يعجب et du pronom relatif ه qui sont tous agglutinés. Dans cette perspective, nous supposons qu'une décomposition en éléments simples du mot complexe pourrait réduire la taille du vocabulaire arabe (comprenant plusieurs centaines de millions de mots) et augmenter les occurrences de chacun des éléments et leurs contextes. Ceci pourrait améliorer la qualité des représentations vectorielles. Dans cet article, notre objectif consiste donc à vérifier et valider cette hypothèse.
Les embeddings pré-entrainés existants représentent un mot arabe sans considération des caractéristiques d'agglutination et de la richesse morphologique de l'arabe. Nous nous focalisons dans ce travail sur la spécificité de la langue arabe et nous construisons des embeddings pour différentes représentations du mot : le token (unité morpho-syntaxique simple, le mot est notamment séparé de la ponctuation), le lemme (forme canonique du token) et le light stemme (suppression des affixes du token). À notre connaissance, c'est le premier travail proposant la construction de tels embeddings. Ces derniers seront prochainement disponibles gratuitement.

Dans cet article, nous nous intéressons à la détection d'opinions par les méthodes d'apprentissage profond pour la langue arabe. Nous effectuons nos experiences sur le corpus *Large-scale Arabic Book Review* (LABR), corpus de critiques de livres en arabe. Nous présentons en section 2 notre méthodologie pour la contruction des embeddings spécifiques à l'arabe. Nous proposons ensuite, en section 3, notre système neuronal basé sur un réseau de neurones convolutifs (CNN) en justifiant le

choix de quelques hyperparamètres suite à une analyse du corpus. Nous présentons et analysons, en section 4, les performances. Finalement, nous concluons et exposons les perspectives de ce travail en section 5.

# 2 Embeddings spécifiques à l'arabe

L'objectif de ce travail consiste à construire des embeddings de mots arabes pré-traités et de les évaluer pour la tâche d'analyse d'opinions [1]. Nous détaillons dans cette section le processus de construction de tels embeddings.

## 2.1 Préparation du corpus d'entraînement des embeddings spécifiques

Nous proposons d'utiliser comme corpus d'entraînement des embeddings spécifiques une fusion de différents corpus arabes existants portant sur la tâche de détection d'opinions ou sur des articles de presse. Nous disposons de quatre de ces corpus. Les trois premiers sont des corpus d'opinion : BRAD (Elnagar & Einea, 2016) regroupant $510k$ critiques sur livres, HARD (Elnagar *et al.*, 2018) constitué de $373K$ commentaires sur films et le corpus d'apprentissage de LABR (Nabil *et al.*, 2014) formé de $23K$ de critiques sur livres. Le dernier corpus *AbuElKhair* (El-Khair, 2016) regroupe $5222k$ articles de presse. Ce corpus sera noté *Global*.

Nous avons nettoyé le corpus fusionné : nous avons supprimé les urls, les mentions, les hashtags, les nombres, les signes de ponctuation et les mots non arabes. Dans ce travail, nous envisageons une classification binaire de commentaires en positifs et négatifs. Les moyens d'accentuation et d'emphase ne nous semblent donc pas très pertinents pour la classification. Nous avons ainsi supprimé le caractère d'allongement (*kashida* ou *tatouil*) et gardé deux occurrences de caractères identiques consécutifs. Nous avons normalisé les caractères أ, آ à un simple alif ا. Nous avons également supprimé les voyelles courtes arabes et les diacritiques (*soukoun* et *chadda*). Ce nettoyage permet de normaliser les différentes possibilités d'écriture de mots en arabe ce qui permet de réduire la taille du vocabulaire et ainsi diminuer le nombre d'hapax.

## 2.2 Construction des embeddings

L'arabe est une langue caractérisée par son agglutination et sa richesse morphologique. L'application d'outils TAL semble être nécessaire pour réduire la fausse diversité du vocabulaire arabe et construire des espaces d'embeddings : de tokens, de lemmes et de light stemmes. Le token est obtenu par un outil qui permet de séparer les clitiques de l'unité morpho-syntaxique simple. La lemmatisation consiste à réduire tous les mots fléchis à leur forme canonique. Le light stemming consiste à supprimer les préfixes et les suffixes fréquemment utilisés avec les mots sans les réduire à leurs racines. Nous avons appliqué un tokeniseur [2], un lemmatiseur [3] et un light stemmer [4] à notre corpus *Global*.

1. Les embeddings proposés peuvent être utilisés dans d'autres tâches NLP tel que le résumé automatique, *etc.*
2. http ://qatsdemo.cloudapp.net/farasa/
3. http ://qatsdemo.cloudapp.net/farasa/
4. https ://github.com/motazsaad/arabic-light-stemming-py

Les différents espaces d'embeddings sont construits en utilisant le modèle word2vec (Mikolov *et al.*, 2013). La version Skip-gram de ce dernier est meilleure pour l'analyse d'opinions en anglais (Kim, 2014) et arabe (Barhoumi *et al.*, 2018). Nous avons donc construit 3 espaces d'embeddings (de dimension 300) dédiés aux 3 différentes unités lexicales (tokens, lemmes et light stemmes) avec le type skip-gram. La table 1 reporte la taille de chaque espace. Pour information, le vocabulaire sous forme de mots est de taille 3 millions.

| | **token** | **lemme** | **light stemme** |
|---|---|---|---|
| **Taille** | 1 980 255 | 946 171 | 1 997 601 |

TABLE 1 – Taille des espaces d'embeddings proposés pour l'arabe.

# 3 Système d'analyse d'opinions pour l'arabe

Les réseaux convolutifs CNN ont prouvé leurs performances dans l'analyse d'opinions pour l'anglais (Kim, 2014) et l'arabe (Barhoumi *et al.*, 2018). Nous avons donc choisi cette architecture pour implémenter notre système et évaluer ses performances pour l'arabe [5]. Nous décrivons dans la suite l'architecture du CNN et nous détaillons le choix de quelques hyperparamètres liés à notre analyse du corpus.

## 3.1 Architecture du CNN

Le CNN prend en entrée une matrice d'*embeddings* de taille fixe et applique une convolution de filtres, dont la taille de la fenêtre est une des valeurs de l'ensemble$\{3, 4, 5\}$, pour extraire de nouveaux attributs à partir de la matrice d'*embeddings*. Puis, un *max_pooling* est appliqué sur la sortie de la couche de convolution dans le but de conserver uniquement les attributs les plus pertinents qui sont concaténés au niveau d'une couche entièrement connectée. Enfin, le CNN applique la fonction *sigmoid* à la couche de sortie pour générer la polarité du document fourni en entrée. Deux polarités sont possibles : positif ou négatif.

## 3.2 Choix d'hyperparamètres

### 3.2.1 Longueur du document

Le réseau convolutif CNN prend comme entrée une matrice de taille fixe (cf. section 3.1). Dans notre cas, l'entrée du CNN sera la matrice représentant un commentaire par l'ensemble de ses mots, chacun d'eux représenté par un embedding. Or, chaque commentaire ne contient pas le même nombre de mots. Il convient donc de définir la taille du document. Pour cela, nous utilisons la formule 1 qui nous permet de déterminer le nombre maximum de mots qui représenteront le document.

$$seuil = moyenne + 2 \times ecart\ type \quad (1)$$

En appliquant cette formule sur le corpus d'apprentissage de LABR, nous calculons une longueur moyenne des commentaires de 64 mots et un écart type de 117,71 mots. Nous obtenons donc un seuil de 300 mots. Ainsi, chaque document sera représenté par une matrice de 300 mots par 300

5. Ce qui nous distingue de (Barhoumi *et al.*, 2018) est les embeddings d'entrée au réseau CNN

composantes (taille de l'embedding d'un mot). Dans le corpus d'apprentissage de LABR, plus de 96% des commentaires contiennent moins de 300 mots. Nous définissons dans la section suivante comment représenter les documents en fonction de leur taille.

### 3.2.2 Padding/Truncating

Lorsque la taille du document est supérieure à celle fixée, il est nécessaire de couper (et ainsi ignorer) les mots supplémentaires : c'est le *truncating*. Lorsque les documents sont plus courts, il est nécessaire de combler la représentation du message par des zéros : c'est le *padding*. Or il existe trois façons de procéder à ce truncating/padding : soit couper/combler sur le début du message (pré), soit sur la fin du message (post), soit de manière égale sur les 2 extrémités.

Pour choisir le protocole padding/truncating le plus adéquat dans notre cas, nous avons procédé à une analyse des mots polarisés contenus dans les documents afin de déterminer quel est le segment qui contient l'information la plus pertinente pour la classification. Dans le cadre de l'analyse d'opinions, cette information regroupe principalement les mots polarisés et les termes de négation qui sont souvent utilisés dans l'expression d'opinions. Pour déterminer la polarité d'un mot, un lexique de mots polarisés a été utilisé. Ce dernier est la fusion de 15 lexiques polarisés existants (arabes ou traduit de l'anglais vers l'arabe). Le lexique résultant contient 51968 mots positifs et 45638 mots négatifs. Pour les termes de négation, une liste prédéfinie regroupe 6 différents termes de négation.

Nous avons effectué des statistiques sur les pourcentages de mots polarisés et de termes de négation pour mesurer l'informativité des segments. Nous avons divisé le document en trois parties et calculé le pourcentage de mots polarisés ou de terme de négation contenu dans chacun des trois segments. Ces statistiques sont reportées dans le tableau 2.

| | | **1$^{ier}$ segment** | **2$^{ieme}$ segment** | **3$^{ieme}$ segment** |
|---|---|---|---|---|
| **TRAIN** | **% mots positifs** | 16,33% | 0,73% | 0,82% |
| | **% mots négatifs** | 7,29% | 0,34% | 0,63% |
| | **% termes de négation** | 0,74% | 0,03% | 0,002% |

TABLE 2 – Informativité des différents segments du corpus Train de LABR

Une première remarque basée sur le tableau 2 porte sur l'informativité du premier tiers du document qui comprend le plus grand pourcentage de mots polarisés. Les deux autres tiers ne sont pas aussi informatifs que le premier. Nous pourrions donc déduire que les internautes expriment explicitement leurs opinions au début du commentaire et ils se justifient par la suite de manière plus factuelle. Le premier tiers de chaque document semble donc contenir de l'information pertinente pour la classification en polarité. Le post-padding/post-truncating semble ainsi être adapté à l'analyse d'opinion du corpus LABR. Si le document comprend plus de 300 mots, la fin de celui-ci sera donc coupé. S'il est plus petit, il sera complété par le vecteur 0 autant que nécessaire.

# 4 Résultats et discussion

## 4.1 Corpus LABR

Pour évaluer notre système, nous avons utilisé le corpus LABR (Nabil *et al.*, 2014) qui contient 63k critiques de livres composées d'un commentaire et d'une note associée (nombre d'étoiles). Nous

nous plaçons dans le cadre d'une classification binaire et regroupons les critiques comme proposé dans (Nabil *et al.*, 2014) : les commentaires associés à une ou deux étoiles composent la classe *négative* et ceux à quatre ou cinq étoiles composent la classe *positive*. Ainsi les commentaires neutres ne sont pas considérés et le corpus utilisé se réduit à un ensemble de 33234 commentaires (84% positifs) pour le corpus d'apprentissage et 8366 pour le corpus de test (85% positifs). Notons que 10% de l'ensemble d'apprentissage est utilisé comme corpus de développement.

## 4.2 Performance du système d'analyse d'opinions

Cette section présente les performances de notre système neuronal. La table 4 rapporte les performances du CNN avec les différents espaces d'embeddings implémentés (section 2.2). Nous remarquons que quel que soit l'espace d'embeddings, l'exactitude est supérieure à 91%. Les performances des trois embeddings implémentés sont équivalentes.

| **LABR** | **unité lexicale** | | | **(Barhoumi *et al.*, 2018)** |
|---|---|---|---|---|
| | **token** | **lemme** | **light stem** | **www-sg** |
| **Dev** | **91,52%** | **91,48%** | 91,36% | 89,82% |
| **Test** | 91,25% | **91,50%** | **91,50%** | 89,34% |

TABLE 3 – Exactitude du CNN sur LABR avec les différents modèles d'embeddings.

Ainsi, les systèmes CNN avec des embeddings d'unités pré-traitées sont plus performants que le CNN appliqué sur des embeddings de mots existants. Nous pensons que ceci revient certainement à une plus grande couverture du vocabulaire et une représentation plus robuste des unités lexicales. Il serait intéressant de valider cette hypothèse dans d'autres tâches de traitement automatique des langues. Nous avons calculé la matrice de confusion du système *CNN_token* (meilleure performance sur le corpus de développement). Le système prédit bien les critiques positives avec 97,68% de précision et 92,46% de rappel. Les commentaires négatifs sont plus difficiles à détecter avec seulement 54,45% de précision et 81,49% de rappel.

Nous avons également comparé notre système avec les résultats des travaux déjà parus sur le corpus LABR. À notre connaissance, les meilleurs résultats obtenus par un système ne se basant pas sur des connaissances *a priori* de type expert sont ceux du système (Barhoumi *et al.*, 2018). Ce dernier a utilisé un CNN et a atteint 89,34% d'exactitude avec des embeddings de mots de type skip-gram pré-entrainés sur des pages web issus de (Soliman *et al.*, 2017).
Dans ce travail, notre système s'appuie sur une architecture CNN similaire à (Barhoumi *et al.*, 2018). Suite à notre analyse du corpus, nous avons réduit la taille du vecteur représentant le document (de 882 à 300) et avons appliqué un processus de post-padding/truncating (au lieu de pré-). Le gain obtenu en ne changeant que la valeur de ces paramètres est de 0.9% en absolu. Le reste du gain est dû à l'utilisation de nos espaces spécifiques d'embeddings. Une des différences repose sur la couverture des espaces d'embeddings sur le corpus LABR. Celle de nos espaces est d'environ 99% alors que celle des embeddings de mots utilisés dans (Barhoumi *et al.*, 2018; Dahou *et al.*, 2016) couvre 81% de LABR. Ceci explique certainement le gain que nous observons.

Nous proposons d'illustrer les différents résultats en fonction de la représentation lexicale via un exemple. La table 4 reporte les différentes représentations lexicales possibles du commentaire تمنيت ان الموضوعَات و الاَحَاديث فيه اكثر /tmnyt An AlmwDw't w AAHdyv fyh Akvr/ (j'aurai aimé

plus de thèmes et de conversations) et dont la classe de référence est positive. Notre architecture CNN utilise la fonction sigmoïde en couche de sortie. Elle associe la classe positive (respectivement négative) si le score de sortie est supérieur (respectivement inférieur) à 0.5. Plus le score est proche de 1 (respectivement 0), plus la prédiction de la classe positive (respectivement négative) est certaine.

| Unité lexicale | Commmentaire | Classe prédite | Score |
|---|---|---|---|
| Token | تمني موضوع احَاديث في أكثر | positive | 0.71 |
| Lemme | تمنَى ان موضوع احَاديث في أكثر | positive | 0.6 |
| Light stem | تمنيت ان موضوع احَاديث فيه اكثر | négative | 0.39 |

TABLE 4 – Exemple de commentaire et sa prédiction selon les différentes unités lexicales .

Pour cet exemple, le CNN prédit correctement la classe positive avec les embeddings de tokens et ceux de lemmes. Nous remarquons que le score obtenu avec les embeddings de tokens (0.71%) est supérieur à celui obtenu avec les embeddings de lemmes (0.6%) et que la classe prédite avec les embeddings de light stems n'est pas correcte (0,39%).

# 5 Conclusion et perspectives

Nous avons souhaité prendre en compte les spécificités de la langue arabe. Ainsi, nous avons proposé et construit 3 espaces d'embeddings prenant en compte l'agglutination et la richesse morphologique de la langue arabe. Nous avons mesuré l'utilité de tels embeddings et nous avons trouvé des performances proches, de l'ordre de 91%. Le meilleur résultat sur le corpus de développement est obtenu par les tokens tandis que sur le corpus de test, il est atteint avec des embeddings de lemmes ou light stemmes. Nous notons une amélioration d'environ 2% par rapport à la baseline (89,34%). À notre connaissance, c'est le premier travail qui contruit et teste les embeddings de différentes représentations lexicales en arabe.

D'autres pistes d'amélioration restent à explorer pour tenir compte du phénomène d'agglutination et de la richesse morphologique de l'arabe. Nous envisageons notamment de tester les embeddings de n-grammes de caractères à la façon de fasttext, les embeddings de mots à base de caractères ou encore les dernières représentations vectorielles *ELMO*. De plus, en nous appuyant sur les travaux de (Yu *et al.*, 2017) où des *embeddings d'opinions* sont construits pour l'anglais, nous souhaitons également étudier la transposition de ces travaux pour l'arabe en nous appuyant sur les lexiques de mots polarisés que nous avons construits.

# Références

ABDULLA N. A., AHMED N. A., SHEHAB M. A., AL-AYYOUB M., AL-KABI M. N. & AL-RIFAI S. (2014a). Towards improving the lexicon-based approach for arabic sentiment analysis. *International Journal of Information Technology and Web Engineering (IJITWE)*, **9**(3), 55–71.

ABDULLA N. A., AL-AYYOUB M. & AL-KABI M. N. (2014b). An extended analytical study of arabic sentiments. *International Journal of Big Data Intelligence 1*, **1**(1-2), 103–113.

AL SALLAB A., HAJJ H., BADARO G., BALY R., EL HAJJ W. & SHABAN K. B. (2015). Deep learning models for sentiment analysis in arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, p. 9–17.

ALMAS Y. & AHMAD K. (2007). A note on extracting 'sentiments' in financial news in english, arabic & urdu. In *The Second Workshop on Computational Approaches to Arabic Script-based Languages*, p. 1–12.

BARHOUMI A., CAMELIN N. & ESTÈVE Y. (2018). Des représentations continues de mots pour l'analyse d'opinions en arabe : une étude qualitative. In *Actes de la conférence Traitement Automatique de la Langue Naturelle, TALN 2018*, p. 215.

BAYOUDHI A., GHORBEL H. & BELGUITH L. H. (2015). Sentiment classification of arabic documents : Experiments with multi-type features and ensemble algorithms. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, p. 196–205.

DAHOU A., XIONG S., ZHOU J., HADDOUD M. H. & DUAN P. (2016). Word embeddings and convolutional neural network for arabic sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, p. 2418–2427.

EL-HALEES A. (2011). Arabic opinion mining using combined. *Proceeding the International Arab Conference On Information Technology*.

EL-KHAIR I. A. (2016). 1.5 billion words arabic corpus. *arXiv preprint arXiv :1611.04033*.

ELNAGAR A. & EINEA O. (2016). Brad 1.0 : Book reviews in arabic dataset. In *Computer Systems and Applications (AICCSA), 2016 IEEE/ACS 13th International Conference of*, p. 1–8 : IEEE.

ELNAGAR A., KHALIFA Y. S. & EINEA A. (2018). Hotel arabic-reviews dataset construction for sentiment analysis applications. In *Intelligent Natural Language Processing : Trends and Applications*, p. 35–52. Springer.

FARRA N., CHALLITA E., ASSI R. A. & HAJJ H. (2010). Sentence-level and document-level sentiment mining for arabic texts. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, p. 1114–1119 : IEEE.

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

IBRAHIM H. S., ABDOU S. M. & GHEITH M. (2015). Sentiment analysis for modern standard arabic and colloquial. *International Journal on Natural Language Computing (IJNLC)*, **4**(2).

KIM Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv :1408.5882*.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, p. 3111–3119.

NABIL M., ALY M. & ATIYA A. (2014). Labr : A large scale arabic sentiment analysis benchmark. *arXiv preprint arXiv :1411.6718*.

PANG B., LEE L. *et al.* (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, **2**(1–2), 1–135.

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

PETERS M. E., NEUMANN M., IYYER M., GARDNER M., CLARK C., LEE K. & ZETTLEMOYER L. (2018). Deep contextualized word representations. In *Proc. of NAACL*.

REFAEE E. & RIESER V. (2016). ilab-edinburgh at semeval-2016 task 7 : A hybrid approach for determining sentiment intensity of arabic twitter phrases. *Proceedings of SemEval-2016*, p. 474–480.

SOLIMAN A. B., EISSA K. & EL-BELTAGY S. R. (2017). Aravec : A set of arabic word embedding models for use in arabic nlp. *Procedia Computer Science*, **117**, 256–265.

YANG K., CAI Y., HUANG D., LI J., ZHOU Z. & LEI X. (2017). An effective hybrid model for opinion mining and sentiment analysis. In *Big Data and Smart Computing (BigComp), 2017 IEEE International Conference on*, p. 465–466 : IEEE.

YU L.-C., WANG J., LAI K. R. & ZHANG X. (2017). Refining word embeddings for sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 534–539.

# Q-learning pour la résolution des anaphores pronominales en langue arabe

Saoussen Mathlouthi Bouzid[1] Chiraz Ben Othmane Zribi[1]
(1) RIADI, ENSI Campus Universitaire de la Manouba, 2010 Manouba, Tunisie
Mathlouthi.saw@gmail.com, Chiraz.zribi@ensi-uma.tn

RÉSUMÉ

La résolution d'anaphores est une tâche fondamentale pour la plupart des applications du TALN. Cette tâche reste un problème difficile qui nécessite plusieurs sources de connaissances et des techniques d'apprentissage efficaces, notamment pour la langue arabe. Cet article présente une nouvelle approche de résolution d'anaphores pronominales dans les textes arabes en se basant sur une méthode d'Apprentissage par Renforcement AR qui utilise l'algorithme Q-learning. Le processus de résolution comporte une étape d'identification des pronoms et des antécédents candidats et une autre de résolution. L'algorithme Q-learning permet d'apprendre dans un environnement dynamique et incertain. Il cherche à optimiser pour chaque pronom anaphorique, une séquence de choix de critères pour évaluer les antécédents et sélectionner le meilleur. Le système de résolution est évalué sur des textes littéraires, des textes journalistiques et des manuels techniques. Le taux de précision atteint jusqu'à 77,14%.

ABSTRACT

**Q-learning for pronominal anaphora resolution in Arabic texts**

Anaphora resolution is a fundamental task for most NLP applications. This task remains a difficult problem that requires several sources of knowledge and effective learning techniques, especially for the Arabic language. This paper presents a novel approach to resolving pronominal anaphora in Arabic texts based on a Reinforcement Learning RL that uses the Q-learning algorithm. The resolution process includes two steps: pronoun and antecedent identification step and the resolution step. The Q-learning algorithm allows learning in a dynamic and uncertain environment. It seeks to optimize for each anaphoric pronoun, a sequence of criteria choice to evaluate the antecedents and look for the best. The resolution system is evaluated on literary texts, journalistic and technical manual texts. Precision rate reaches until 77.14%.

MOTS-CLÉS : résolution d'anaphores, apprentissage par renforcement, Q-learning, critères morphosyntaxiques, arabe.

KEYWORDS: anaphora resolution, reinforcement learning, Q-learning, morpho-syntactic criteria, Arabic.

# 1 Introduction

Dans les discours en langue naturelle, plusieurs moyens différents permettent la référence à des objets ou des entités. Dans certaines situations linguistiques, les répétitions des Groupes Nominaux (GNs), peuvent être réduites au pronom afin d'éviter la redondance et la lourdeur ; c'est le

phénomène d'anaphore. Ce phénomène linguistique joue un rôle important dans la construction du sens. Il met en œuvre les différentes possibilités de reprise d'un élément qui peut paraître et réapparaître dans un texte. Chaque expression anaphorique dépend d'une autre expression (appelée référence ou antécédent) qui doit être trouvée dans la partie antécédente (ou parfois suivante) du texte. Dans notre travail nous avons traité les pronoms personnels sujets et objets, les pronoms démonstratifs et les pronoms relatifs. En langue arabe, les anaphores pronominales sont variés et denses ; nos statistiques réalisés sur des textes littéraires (comme l'exemple 1) montrent que 18,2% des mots sont des pronoms. En outre, la langue arabe est une langue riche morphologiquement et présente plusieurs spécificités qui rendent la tâche de résolution plus ardue, comme expliqué dans (Mathlouthi et al., 2016).

(1) فكه عينيك **بتلك** البسط الخضراء **التي** نسجتـ**ها** يد الطبيعة نفسـ**ها فتلك هي** السعادة بعينـ**ها**

Jouissez vos yeux de ces vallées verdoyantes qui ont été créées par la nature elle-même, c'est tout le bonheur

La tâche de résolution d'anaphores pronominales (RAP) comporte deux étapes principales : une étape préliminaire pour l'identification des pronoms et une étape de résolution. Pour l'étape de résolution, nous considérons un ensemble de critères morphosyntaxiques nécessaires pour choisir le meilleur antécédent de chaque pronom parmi la liste des antécédents possibles. A cet effet, nous proposons une approche d'apprentissage par renforcement utilisant ces critères morphosyntaxiques comme étant l'ensemble des actions à choisir pour juger le bon candidat. Le choix de l'apprentissage par renforcement est motivé par les raisons suivantes :

- En langue arabe, le manque des données de taille importante et étiquetées avec des liens anaphoriques rend parfois l'utilisation de l'apprentissage complètement supervisé assez difficile.
- L'environnement du système de résolution est dynamique, car d'une part la liste des antécédents est limitée à une fenêtre de mots, et d'autre part, les critères linguistiques et leurs pertinences peuvent changer selon le pronom et le style du texte traité.
- Le système de RAP cherche à optimiser une séquence de décisions (choix des critères) afin de trouver le meilleur antécédent candidat.

L'apprentissage par renforcement est une technique utilisée pour permettre à un agent de connaître son environnement et de savoir quand explorer et quand exploiter pour prendre une bonne décision. L'algorithme Q-learning est l'une des techniques d'apprentissage par renforcement les plus utilisées. Il réalise l'équilibre entre les processus d'exploration et d'exploitation. L'environnement de notre système de RAP est modélisé par un processus de décision de Markov (PDM) qui permet de représenter, pour chaque pronom et ses antécédents candidats, le choix des combinaisons de critères possibles.

Le présent article est composé de quatre sections. Dans la section 2, nous menons une étude comparative de l'état de l'art entre les différents travaux existants. Nous détaillons les étapes de notre approche de RAP, dans la section 3, et nous expliquons l'utilité de la méthode d'apprentissage par renforcement AR dans la tâche de résolution. Enfin, nous présentons notre corpus de test, les résultats des expérimentations réalisées et leurs comparaisons aux résultats d'un travail similaire.

# 2 Les travaux antérieurs

La tâche de résolution d'anaphore a été le sujet de recherche de plusieurs travaux en TALN. Les travaux peuvent être classés en quatre types d'approche : les approches à base de règles, les

approches statistiques, les approches à base d'apprentissage et les approches hybrides. Les approches à base de règles linguistiques exploitent plusieurs sources de connaissances telles que les travaux Lappin & Leass (1994), Mitkov (1998) et Schmolz et al. (2012) pour l'anglais. Le travail Mitkov (1998) a été adapté à la langue arabe dans Mitkov et al. (1998). Néanmoins, les sources de connaissances linguistiques restent insuffisantes pour résoudre la complexité de la tâche surtout pour certaines langues naturelles ayant une structure linguistique très variée comme l'arabe. En fait, les règles linguistiques sont incapables à elles seules de résoudre des ambiguïtés sémantiques voire même pragmatiques. Certains travaux se sont reposés sur des méthodes à base de calcul statistiques comme le travail de Seminck & Amsili (2017). D'autres travaux ont utilisé des méthodes d'apprentissage afin de couvrir les insuffisances des règles linguistiques. La plupart des travaux dans cette classe ont considéré la résolution comme un problème de classification et ils ont exploité les vecteurs caractéristiques des paires pronoms-antécédents. Plusieurs d'entre eux ont utilisé l'apprentissage supervisé en exploitant l'entraînement des données étiquetées comme le travail d'Aone & Bennett (1996) pour le japonais, Li et al. (2011) pour l'anglais et Aktas et al. (2018) pour la langue allemande. Toutefois, l'apprentissage supervisé nécessite des sources de données étiquetées avec des liens anaphoriques de taille importante, ce qui est parfois coûteux et difficile à réaliser pour certaines langues. Les approches basées sur l'apprentissage non-supervisé, comme le travail de Charniak & Elsner (2009), sont moins nombreux. Ainsi, l'apprentissage permet de couvrir l'incertitude du domaine linguistique et les divers niveaux d'ambiguïtés dans les langues naturelles. Pour les approches hybrides, les auteurs ont combiné les règles linguistiques et les techniques d'apprentissage dans une seule représentation pour tirer profit de leurs avantages respectifs et pour que l'une couvre les insuffisances de l'autre. Parmi les travaux qui ont opté pour ce type d'approche sont : Weissenbacher & Nazarenko (2007), Kamune & Agrawal (2015) pour l'anglais, Abolohom & Omar (2015) et Hammami (2016) pour l'arabe. Nous déduisons que les résultats des performances pour ce dernier type d'approches, particulièrement pour certaines langues comme la langue arabe, restent toujours insuffisants et nécessitent beaucoup plus d'efforts.

# 3 Approche d'apprentissage par renforcement basée sur Q-learning

L'objectif de la RAP consiste à chercher le meilleur antécédent du pronom anaphorique parmi la liste des antécédents candidats. Le système de résolution comporte deux étapes principales, à savoir: l'identification et la résolution. Les pronoms sont identifiés par leurs valeurs grammaticales puis ils sont filtrés en utilisant une approche à base de règles afin d'éliminer les pronoms non-référentiels. Nous nous limitons dans ce qui suit à la présentation de l'étape de résolution, l'étape d'identification des pronoms qui inclus le filtrage des pronoms non-référentiels est décrite en détail dans (Mathlouthi et al., 2016).

## 3.1 Approche de résolution des anaphores pronominales à base d'apprentissage par renforcement

Notre système de résolution cherche le meilleur antécédent de chaque pronom, en utilisant des critères linguistiques qui favorisent certains candidats par rapport à d'autres. La combinaison de critères qui permet de juger le meilleur candidat de chaque pronom est inconnue au préalable et elle varie surtout avec le contexte du pronom. Nous avons ainsi opté pour une approche à base d'apprentissage par renforcement car ce dernier constitue une méthode efficace pour apprendre dans un environnement incertain et dynamique. L'environnement de notre système comporte le pronom,

ses informations morphosyntaxiques et la liste des critères linguistiques. L'agent système de résolution d'anaphores se charge d'apprendre par lui-même tout en interagissant avec son environnement. Il renforce les actions qui s'avèrent être les meilleures, et ce, dans le but de maximiser les récompenses obtenues à l'issue de chaque action. Dans l'apprentissage automatique, l'environnement est modélisé comme un PDM. Le Q-learning est introduit pour affecter la décision d'un agent afin d'explorer plus et d'améliorer sa décision. Il réalise l'équilibre entre les processus d'exploration et d'exploitation. La figure 1 décrit le processus de RAP. Le processus de résolution permet de parcourir le texte traité et d'identifier chaque fois le pronom et la liste d'antécédents correspondante. L'algorithme Q-learning utilise une matrice de récompense R et interagit avec son environnement contenant le contexte du pronom et une liste de critères. Cette matrice R est initialisée lors d'une phase de pré-apprentissage qui utilise quelques textes étiquetés. Cependant, l'algorithme Q-learning permet de trouver la combinaison de critères optimale, utilisée pour évaluer les antécédents et choisir le meilleur d'entre eux (Figure 1).

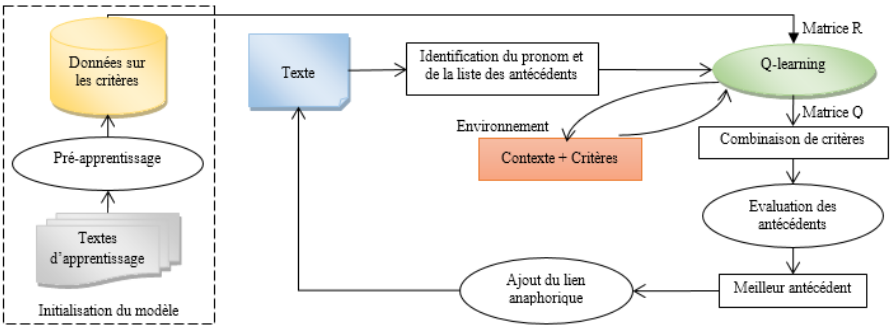



FIGURE 1 : Processus de résolution des pronoms basé sur Q-learning

### 3.1.1 Critères de résolution

Les critères d'évaluation des antécédents sont plus ou moins efficaces. Ils représentent des préférences et pas des facteurs absolus. Leurs pertinences dépendent du contexte du pronom anaphorique et même du style du texte et se sont des estimations de comptages réalisés sur quelques textes étiquetés anaphoriquement. Les critères « Définition », « Thème », « Distance », « Tête de paragraphe », « Nom propre » et « Répétition » sont décrites dans le travail Mathlouthi et al. (2017). Le critère « Définition » favorise les GNs définis à ceux non définis. Le critère « Thème » considère que les sujets des phrases sont préférés par rapport aux autres candidats. Le critère « Distance » considère que les candidats les plus proches sont les plus saillants. Le critère « Tête de paragraphe » favorise le candidat sujet de la première phrase du paragraphe qui reste souvent le centre d'intérêt dans un paragraphe. « Nom propre » est un critère qui favorise des éléments de discours importants. Le critère « Répétition » favorise les candidats dont les lemmes se répètent plusieurs fois dans le texte. En plus, nous avons ajouté un nouveau critère « Antécédent pronom précédent », ce critère privilégie le candidat qui a été déjà choisi comme antécédent pour le pronom précédent. En fait, dans les textes arabes littéraires, ce type de critère est bien vérifié (environ 35% des antécédents vérifient ce critère). Dans l'exemple (2) l'anaphore vérifie les critères « Définition » et « Thème ». Dans certains cas le critère peut pénaliser un antécédent correct, mais cette erreur peut être propagée ou corrigée par un autre critère. De ce fait, les critères ne participent pas nécessairement ensemble dans la résolution d'un pronom particulier. Toutefois, les critères proposés ne sont pas définitifs, ils dépendent des textes traités et peuvent être augmentés par d'autres.

(2) خرج الملك من اليمن غازيا في جيش (…).وعنت **لسلطته** مصر وافريقية

Le roi a quitté le Yémen pour la guerre avec une armée (…). L'Egypte et l'Afrique ont été soumis à **son** autorité

### 3.1.2 Approche Q-learning

Notre système de résolution est modélisé par un PDM (état, action, transition, récompense). L'ensemble des états comportent l'état initial, les états intermédiaires représentant toutes les combinaisons de critères possibles et l'état final. L'état initial $S_I$ du PDM contient des informations sur le pronom Pr, mais la combinaison des critères (CC) est encore inconnue. Les actions possibles sont les choix de critères. Chaque transition d'un état $S_i$ à un autre $S_j$ a une valeur de récompense associée $r_{ij}$. L'état final $S_F$ contient la séquence d'actions optimale qui revient à la meilleure combinaison de critères. Chaque état $S_i$ peut passer directement à l'état final $S_F$ avec une récompense $r_{iF}$.

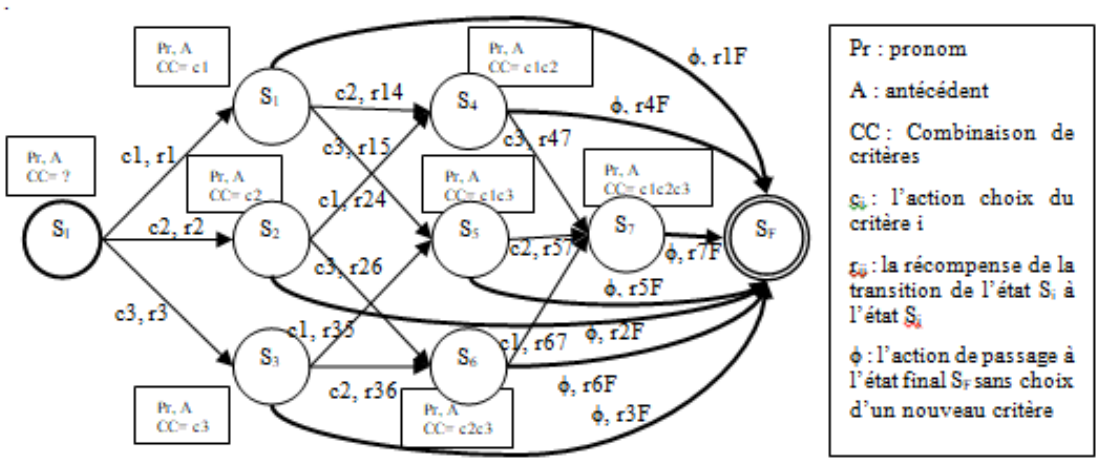



FIGURE 2 : Modélisation PDM pour la résolution des pronoms

La figure 2 montre un exemple de représentation PDM pour le cas de 3 critères. La récompense du choix de l'action $c_x$ sur l'état $S_i$ est calculée par la probabilité P de $S_i$ sachant Pr et $c_x$. Le calcul de la probabilité se base sur la fréquence de la participation de la combinaison de critères dans la résolution des liens anaphoriques sachant le contexte de Pr. Pour le cas du passage direct d'un état $S_i$ de contexte $\{c_y\}$ à l'état final, la récompense $r_{iF}$ est la probabilité pour que l'ensemble de critères $\{c_y\}$ participent seuls dans la bonne résolution.

L'algorithme Q-learning utilise deux matrices Q et R. La matrice R est une matrice à deux dimensions ; les lignes représentent l'ensemble des états et les colonnes sont les actions. Les actions sont les critères $c_x$ et l'action finale $\phi$ qui permet de passer directement à l'état final. Les contextes des états contiennent toutes les combinaisons de critères possibles. Pour chaque état, il y a des actions possibles (leurs récompenses sont les probabilités P décrites précédemment) et d'autres interdites (leurs récompenses valent -1). La matrice Q est initialisée à 0 et elle est mis à jour aux cours des expériences réalisées par l'agent en utilisant la matrice de récompense R. Grâce à cette matrice Q, l'agent met à jour les traces de ses décisions prises dans le passé. Il apprend par expérience et explore d'un état à l'autre jusqu'à atteindre l'objectif. Au niveau de la matrice Q finale, l'ensemble des actions optimales correspond à la meilleure combinaison de critères capable d'évaluer les antécédents du pronom en question. L'algorithme Q-learning procède comme suit :

Début
1. Initialiser les paramètres alpha $\alpha$, gamma $\gamma$ et le nombre d'épisode maximale $E_{max}$,
2. Définir les récompenses de l'environnement dans la matrice R.
3. Initialiser la matrice Q à 0.

4. Pour chaque épisode :
   Sélectionner un état initial aléatoire $S_i$.
   Tant que l'état final est non atteint.
   a. Sélectionner l'une des actions possibles c pour l'état actuel.
   b. Obtenir, à partir de la matrice Q, la valeur Q maximale pour l'état suivant $S_{i+1}$ en fonction de toutes les actions possibles.
   c. Mettre à jour la matrice Q:
   $Q(S_i, c) \leftarrow Q(S_i, c) + \alpha * [R(S_i, c) + \gamma * \text{Max}_{c \text{ dans } C}(Q(S_{i+1}, c))]$
   d. Définir l'état suivant comme état actuel.
   Fin Tant que
   Fin Pour
5. Parcourir la matrice Q finale ; à partir de l'état initial, trouver les actions avec les valeurs de Q les plus élevées jusqu'à atteindre l'état final.

Fin

Le cœur de l'algorithme est une mise à jour de la fonction de valeur (action-état) représentée par $Q(S_i,c)$. A chaque choix d'un critère c, l'agent observe la récompense $R(S_i, c)$ et le nouvel état $S_{i+1}$ et met à jour la matrice Q. Les paramètres $\alpha$ et $\gamma$ ont une plage de 0 à 1 ; $\alpha$ est un facteur d'apprentissage, il contrôle le taux de mise à jour, $\gamma$ est un facteur d'actualisation pour modérer l'effet des récompenses futures.

### 3.2.3. Evaluation des antécédents

L'algorithme Q-learning permet de sélectionner la meilleure combinaison de critères CC pour chaque pronom Pr. D'autre part, chaque antécédent candidat A vérifie un ensemble de critères. En fait, notre but est de donner un score à chaque antécédent pour pouvoir l'évaluer. Le score d'un antécédent dépend de la pertinence des critères CC vérifiés. Si l'antécédent $A_i$ vérifie le critère c ($Verif(A_i, c)=1$) alors son score augmente en ajoutant la pertinence sinon son score diminue ($Verif(A_i, c)=-1$). Les scores d'évaluations permettent de juger le meilleur antécédent. Le score d'évaluation calculé pour chaque antécédent est décrit par la formule (3).

$$score_{Eval} = \sum_{\forall\, c \in CC} (Verif(A, c) * Pertinence(c)) \quad (3)$$

# 4 Expérimentations et résultats

Afin d'évaluer la performance de l'approche, nous avons réalisé plusieurs expérimentations sur un corpus de textes variés. Le corpus est composé de textes littéraires présentés dans le livre scolaire de la 8ème année de base de l'enseignement tunisien, et aussi des textes journalistiques et des manuels techniques extraits du web. Ce corpus contient 4201 mots et 436 pronoms dont 409 sont référentiels. La phase de pré-apprentissage utilise des textes d'entraînement contenant 5196 mots et 638 pronoms. En fait, le simulateur implémenté prend comme paramètre d'entré des textes d'entraînement de même type que le texte traité. Notons que le corpus d'entraînement est utilisé juste pour initialiser le modèle mais n'intervient pas lors de la phase d'apprentissage par renforcement. Notre système a réussi à détecter tous les pronoms anaphoriques et à les identifier selon leurs types. Il couvre toutes les anaphores considérées dans la résolution et génère une liste de candidats non vide pour la plupart des pronoms. Afin de mieux apprécier l'approche proposée, nous avons implémenté et adapté l'approche robuste de Mitkov et al. (1998) (qui a pu atteindre des bons résultats) et nous l'avons évaluée sur nos textes de test afin de comparer les résultats. Nous avons choisi l'ensemble des critères de préférences proposés par Mitkov qui s'adaptent à la langue arabe à savoir , « Définition », « Thème », « Distance » et « Tête de paragraphe » et nous l'avons augmenté

par nos critères « Nom propre », « Répétition » et « Antécédent pronom précédent ». Nous avons utilisé les pertinences (-1, 0, 1, 2) que Mitkov a attribué à ces critères. La Table 1 présente les résultats d'évaluation de notre approche sur des textes littéraires, des manuels techniques et des textes journalistiques. L'approche de Mitkov a été testée également sur les mêmes textes.

| Textes | | Textes littéraires | Textes manuels techniques | Textes journalistiques |
|---|---|---|---|---|
| Taille | Nombre mots | 1615 | 1714 | 872 |
| | Nombre pronoms | 218 | 136 | 82 |
| Approche Q-learning | | 70,11% | 72,8% | **77,14%** |
| Approche Mitkov | | 64,67% | 53,04% | 50% |

TABLE 1 : Précisions de l'approche Q-learning et celle de Mitkov

En comparant les résultats de notre approche à ceux de Mitkov, nous avons déduit l'efficacité de l'approche Q-learning. Cette dernière donne des résultats meilleurs que celle de Mitkov pour tous les types de textes traités. La précision pour que l'antécédent ait la position n°1 atteint un taux de 77,14%, nous considérons que ces résultats sont encourageants. Pour mettre en évidence une amélioration possible des résultats, nous avons présenté dans la figure 3 la précision selon la position de l'antécédent correct. Si l'antécédent figure parmi les deux premiers candidats (respectivement les trois premiers candidats), alors la précision peut atteindre 84,29% pour les textes journalistiques (respectivement 87,7% pour les textes manuels techniques).

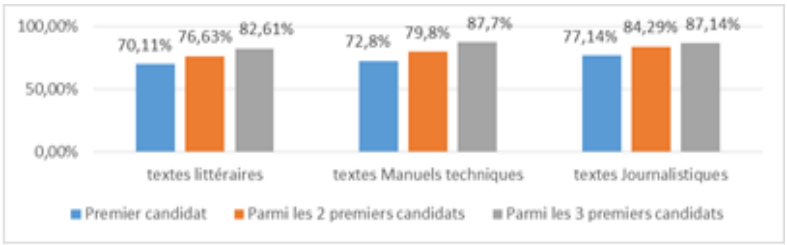



FIGURE 3 : Position de l'antécédent correct parmi la liste des antécédents candidats

# 5 Conclusion

Cet article présente une nouvelle approche de résolution des anaphores pronominales en langue arabe. L'approche basée sur l'apprentissage par renforcement utilise la méthode Q-learning et exploite un ensemble de critères linguistiques. En effet, notre système est modélisé par un PDM qui représente la séquence des critères possibles. L'algorithme Q-learning cherche la combinaison de critères avec les valeurs de récompense les plus élevées. Cette combinaison de critères est utilisée pour évaluer les antécédents et en sélectionner le meilleur. Comme futurs travaux, nous visons agrandir notre corpus de textes journalistiques et de manuels technique afin de réaliser plus d'expérimentations. Nous projetons également d'ajouter des critères sémantiques capables d'améliorer les résultats.

# Références

Mitkov R. (1998). Robust pronoun resolution with limited knowledge. Montreal, Canada: Proceedings of the 18.th International Conference on Computational Linguistics (COLING' 98)/A CL' 98.

Mitkov R., Belguith L., Stys M. (1998). Multilingual robust anaphora resolution, Granada, Spain.

Schmolz H., Coquil D., Döller M. (2012). In-Depth Analysis of Anaphora Resolution Requirements. 2012 23rd International Workshop on Database and Expert Systems Applications. Vienna, Austria.

Seminck O., Amsili P. (2017). A Computational Model of Human Preferences for Pronoun Resolution. Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics, pages 53–63, Valencia, Spain.

Ashima A., Mohana B. (2016). Improving Anaphora Resolution by Resolving Gender and Number Agreement in Hindi Language using Rule based Approach. Indian Journal of Science and Technology, Vol 9(32), august 2016.

Aone C., Bennet S. (1995). Applying machine learning to anaphora resolution. International Joint Conference on Artificial Intelligence IJCAI 1995: Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing pp 302-314, 1995.

Elghamry K., Al-Sabbagh R., El-Zeiny N. Arabic Anaphora Resolution Using Web as Corpus. Proceedings of the seventh conference on language engineering, pp. 1-18. Cairo, Egypt, 2007.

Weissenbacher D., Nazarenko A. (2007). A bayesian classifier for the recognition of the impersonal occurrences of the it pronoun. In Proceedings of DAARC'07, 2007. 29, 32, 43, 72, 120.

Hammami S. (2016). La résolution automatique des anaphores pronominales pour la langue arabe, thèse de doctorat, Université de Sfax, Faculté des Sciences Economiques et de Gestion, Sfax, Tunisie.

Mathlouthi S., Ben Fraj Trabelsi F., Ben Othmane Zribi C. (2016). A Novel Approach Based on Reinforcement Learning for Anaphora Resolution. 28th IBIMA Conference, November 2016.

Mathlouthi S., Ben Fraj F., Ben Othmane C. (2017). How to combine salience factors for Arabic Pronoun Anaphora Resolution. 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications.

Lappin S., Leass H. J. (1994). An Algorithm for Pronominal Anaphora Resolution. London, Computational Linguistics, 20(4), 535-561.

Li D., Miller T., Schuler W. (2011). A pronoun anaphora resolution system based on factorial hidden markov models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, June 19-24*, page 1169–1178, 2011. 42, 115.

Aktas B., Scheffler T., Stede M. (2018). Anaphora Resolution for Twitter Conversations: An Exploratory Study. Proceedings of the Workshop on Computational Models of Reference, Anaphora and Coreference, pages 1–10. New Orleans, Louisiana, June 6, 2018.

Charniak E., Elsner M. (2009). EM works for pronoun anaphora resolution. In Proceedings of EACL, pp. 48-156.

Kamune K., Agrawal A. (2015). Hybrid Approach to Pronominal Anaphora Resolution in English Newspaper Text. International Journal of Intelligent Systems and Applications, 02: pp. 56-64. Published Online January 2015 in MECS. DOI: 10.5815/ijisa.2015.02.08.

Abolohom A., Omar N. (2015). A Hybrid Approach to Pronominal Anaphora Resolution in Arabic, Journal of Computer Sciences 11(5): 764-71 DOI: 10.3844/jcssp.2015.764.771.

# Représentation sémantique distributionnelle et alignement de conversations par chat

Tom Bourgeade Philippe Muller
IRIT, Université de Toulouse
tom.bourgeade@irit.fr, philippe.muller@irit.fr

RÉSUMÉ

Les mesures de similarité textuelle ont une place importante en TAL, du fait de leurs nombreuses applications, en recherche d'information et en classification notamment. En revanche, le dialogue fait moins l'objet d'attention sur cette question. Nous nous intéressons ici à la production d'une similarité dans le contexte d'un corpus de conversations par *chat* à l'aide de méthodes non-supervisées, exploitant à différents niveaux la notion de sémantique distributionnelle, sous forme d'*embeddings*. Dans un même temps, pour enrichir la mesure, et permettre une meilleure interprétation des résultats, nous établissons des alignements explicites des tours de parole dans les conversations, en exploitant la distance de Wasserstein, qui permet de prendre en compte leur dimension structurelle. Enfin, nous évaluons notre approche à l'aide d'une tâche externe sur la petite partie annotée du corpus, et observons qu'elle donne de meilleurs résultats qu'une variante plus naïve à base de moyennes.

ABSTRACT

**Distributional semantic representation and alignment of online chat conversations**

Textual similarity measures have an important place in NLP, because of their many applications, particularly in information retrieval and classification. However, dialog receives less attention on this issue. We are interested here in the production of a similarity measure in the context of a corpus of online chat conversations using non-supervised methods, exploiting at different levels the notion of distributional semantics, in the form of embeddings. At the same time, to enrich the measure, and to allow a better interpretation of the results, we establish explicit alignments of speaker turns in these conversations, using Wasserstein's distance, which allows us to take into account their structural dimension. Finally, we evaluate our approach using an external task on the small annotated part of the corpus, and observe that it yields better results than a naive variant based on averages.

MOTS-CLÉS : similarité textuelle, analyse de conversations, représentations sémantiques, sémantique distributionnelle, distance de Wasserstein.

KEYWORDS: textual similarity, dialog analysis, semantic representation, distributional semantics, Wasserstein distance.

# 1 Introduction

Les problèmes de similarité textuelle ont connu un essor rapide en TAL, en passant de questions au niveau lexical avec des relations de proximité sémantique puis rapidement au niveau phrastique, avec la question de la paraphrase ou l'inférence (Cer *et al.*, 2017), et même au-delà sur la similarité de passages textuels plus grands (Kusner *et al.*, 2015; Le & Mikolov, 2014). Les productions qui ne sont pas dans un cadre mono-locuteur sont relativement négligées, à l'exception de l'*embedding* de tours de parole isolés (Yang *et al.*, 2018), voire du cas particulier de la similarité de questions (Nakov *et al.*,

2017), mais ces modèles ne prennent pas non plus en compte la structure au-delà de la phrase. Nous nous intéressons ici au cas de la similarité de conversation, où l'échange est structuré en tours de parole produits par deux interlocuteurs, dans le cadre de dialogues orientés par une tâche, où chacun des interlocuteurs joue un rôle différent. C'est la similarité globale de conversations entières qui est ciblée. La popularité croissante d'échanges écrits sous la forme de dialogues textuels (*chat*, forum, micro-blogging) soulève des questions intéressantes de mises en rapport de conversations similaires, avec des applications directes à la fouille de conversation, souvent motivées par des problématiques de gestion de la relation client (*Customer Relationship Management*), qui nous sert de cas d'étude.

Dans ce contexte, les applications potentielles d'une définition de similarité de conversation sont nombreuses : par exemple, la possibilité de construire des marches à suivre "types", en regroupant des conversations portant sur un problème technique similaire et/ou dont les étapes de résolution suivies sont similaires ; ou bien encore, la possibilité d'effectuer une recherche rapide d'une conversation similaire pendant qu'une autre se déroule, afin de guider le conseiller, sans passer par des méthodes de recherche classique, telle que la recherche par mots-clés, qui ne renvoient pas toujours un petit nombre de résultats pertinents. Pour établir cette mesure de similarité conversationnelle, on ne s'intéresse pas ici qu'à la dimension sémantique des messages envoyés par chacun des participants, mais également à la structure globale de la conversation elle-même. De plus, le corpus utilisé ici ne disposant que d'une très faible proportion de données annotées (<1%), on n'emploie ici que des méthodes non-supervisées.

Un autre aspect important dans ce cadre d'assistance est l'explicitation des liens entre conversations qui permet de décomposer la similitude des cas rencontrés dans une tâche de support, et de comprendre la pertinence des conversations mises en relation.

La contribution du travail porte donc ici sur les aspects suivants : une définition de similarité entre conversations, avec un alignement explicite des parties de conversations qui sous-tendent cette similarité, et une méthode non-supervisée pour calculer les représentations utilisées pour calculer la similarité. Pour montrer l'intérêt de l'alignement, nous comparons l'utilisation de cette méthode avec d'autres plus simples pour prédire le résultat d'une tâche de classification annexe par plus proches voisins.

# 2 Représentation de conversations

Pour cette tâche, on a choisi une approche plus modulaire que monolithique, afin de pouvoir évaluer indépendamment différentes méthodes et architectures existantes. L'objectif final étant de pouvoir comparer et mesurer la similarité entre des conversations, on choisit de définir une conversation, dans le contexte du corpus utilisé, comme étant constituée d'une séquence de messages accompagnés de l'identité de leurs auteurs respectifs. De ce fait, la tâche s'organise assez naturellement autour de deux sous-objectifs principaux : dans un premier temps, on va chercher à construire des représentations sémantiques numériques, sous la forme de vecteurs dans un espace d'*embedding*, pour chacun des tours de parole qui font partie de la conversation. Puis, dans un second temps, en s'appuyant sur ces représentations, on va construire une représentation de la conversation qui prenne en compte sa structure, et dont on pourra mesurer une distance avec les représentations d'autres conversations, que l'on assimilera à une mesure de similarité.

En pratique ici, on a choisi d'utiliser un modèle encodeur-décodeur type *seq2seq* (Sutskever *et al.*, 2014) pour construire des représentations sémantiques vectorielles des tours de parole. En particulier, on a choisi une implémentation du modèle non-supervisé *Skip-Thought Vectors* (Kiros *et al.*, 2015), en travaillant sur des séquences de *word embeddings* produit par un modèle *FastText* (Bojanowski

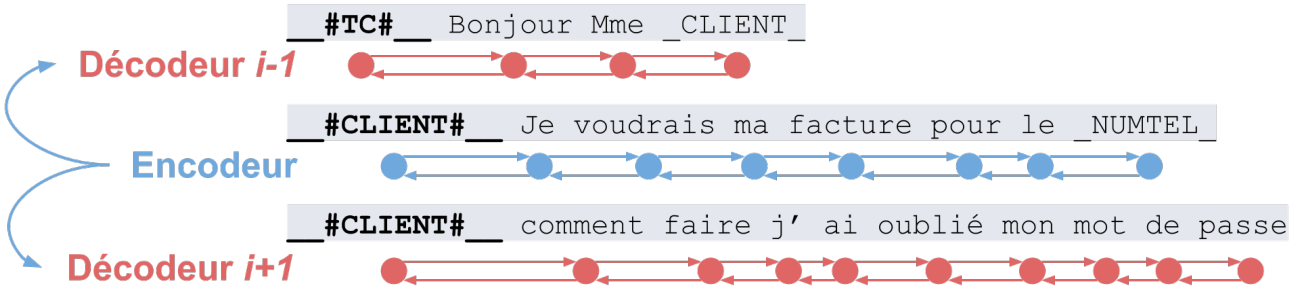



FIGURE 1 – Schéma du modèle *Skip-Thought* utilisé. Le modèle encodeur produit une représentation du $i$-ème message (concaténation des états cachés finaux des deux directions du modèle récurrent) qui est ensuite utilisée pour conditionner deux décodeurs récurrents respectivement sur les $(i-1)$ et $(i+1)$-èmes messages, en espérant ainsi forcer cette représentation à capturer des informations sémantiques conversationnelles.

*et al.*, 2017), tout deux entraînés sur les données non-annotées du corpus. Due à la nature du corpus, il est nécessaire de prendre en compte un certain nombre de contraintes sur le contenu textuel des conversations : en particulier, on a choisi un modèle *FastText* car celui-ci peut obtenir des *embeddings* pour des mots hors-vocabulaire à l'aide d'*embeddings* de *n-grammes*, ce qui nous permet de mieux gérer la présence de fautes de frappe et d'orthographe. Le modèle *Skip-Thought Vectors* produit des encodages de séquences de mots, en prenant en compte leurs contextes immédiats durant l'apprentissage (ici, le tour de parole directement suivant et précédent, voir figure 1), ceci permettant en théorie d'enrichir les informations sémantiques extraites par cet encodeur. En sortie, on obtient donc des représentations vectorielles des messages d'une conversation résultants de la concaténation des états cachés finaux du modèle récurrent bidirectionnel, dont les hyper-paramètres utilisés dans nos expériences sont donnés en table 1.

Une fois les messages des conversations encodés, et afin de prendre en compte la dimension structurelle de celles-ci, on a choisi de s'inspirer de la mesure *Word Mover's Distance (WMD)* présentée dans Kusner *et al.* (2015), elle-même proche de Wan (2007) : cette méthode consiste normalement à employer la distance de Wasserstein, qui mesure le "coût" nécessaire pour transformer une distribution de probabilité en une autre, sur des phrases sous la forme de séquences de *words embeddings*, en résolvant le problème de transport optimal *Earth Mover's Distance (EMD)* associé : on assimile chaque vecteur-mot de la phrase $A$ à un tas de terre et chaque vecteur-mot de la phrase $B$ à un trou, tous de même capacité et définis dans le même espace muni d'une mesure de distance (ou coût) donnée, le but étant de trouver l'ensemble de déplacements de coût total minimum qui permette de remplir tous les trous (ou d'épuiser tous les tas de terre ainsi). Cette approche fournit à la fois une mesure de distance, assimilable à une mesure de similarité sémantique ici, et un alignement optimal (au sens du problème *EMD*) entre deux phrases données. On peut adapter cette méthode, à un niveau d'abstraction plus élevé, en travaillant sur des séquences d'*embeddings* de phrases aux seins de conversations, comme c'est notre cas ici. La résolution d'instances du problème *EMD* peut se faire à l'aide d'un solveur dédié (tel que *POT*[1]), et ne nécessite que de calculer la matrice des coûts entre les vecteurs, à l'aide de la similarité cosinus ou d'une autre norme vectorielle par exemple. L'avantage principal de cette approche est qu'en plus d'obtenir une mesure de similarité conversationnelle structurelle, on obtient également un alignement optimal des messages des conversations traitées, qui peut être utilisé comme entrée pour une autre tâche, ou pour permettre une interprétation qualitative des résultats : en effet on

1. Source : `https://github.com/rflamary/POT`

pourra alors observer comment se transpose l'enchaînement d'actes de dialogue d'une conversation à une autre qui lui est similaire (d'après cette mesure) et ainsi pouvoir vérifier quelles parties de celles-ci ont été jugées similaires, et de quelle manière elle le sont (dans la forme globale que prennent certaines sections, ou dans la thématique d'un ou plusieurs messages en particulier, par exemple). Avec cette approche, une conversation sera donc représentée par la distribution des messages qui la composent dans leur espace d'*embedding*, ceci s'opposant aux méthodes qui agrège ces vecteurs en une représentation simple, qui ont pour avantage d'être plus économes en mémoire et en temps de calcul (pour effectuer des mesures de similarité), mais ont pour désavantage de détruire au moins en partie l'information structurelle de la conversation.

# 3 Données utilisées

Les données utilisées ici proviennent d'un corpus constitué de journaux de conversations par *chat*, entre des téléconseillers et des clients, provenant de la plateforme d'assistance technique et commerciale en-ligne de l'opérateur téléphonique Orange [2]. Chaque conversation est constituée d'une suite de messages horodatés et munis d'un marqueur d'identité anonymisé, avec éventuellement des méta-données associées au contexte technique de la conversation. Une petite partie du corpus a été annotée par les utilisateurs télé-conseillers de la plateforme, avec des labels indiquant l'état de résolution du problème technique du client à la fin de la conversation (par exemple : `PbTechResolu`, si le problème du client a été résolu en fin de dialogue, `InfoFournie` s'il s'agissait d'une question posée par le client et qu'une réponse satisfaisante a été fournie, `SuspenduClient` si le client à mis fin à la conversation abruptement, etc.).

La nature de ce corpus implique quelques spécificités par rapport à des corpus conversationnels plus classiques. Ne s'agissant pas de transcriptions de dialogues oraux, le contenu textuel est fourni tel qu'il a été saisi par les interlocuteurs du *chat*, ce qui implique la présence en abondance de fautes de frappes, de grammaire, d'orthographe, et de problèmes de structure dans les messages (atténuées ici par l'emploi de *FastText*, moins sensible aux petites différences morphologiques lors de la production de *word embeddings*). De plus, le *chat* en-ligne est un medium dans lequel on trouve des phénomènes linguistiques particuliers, dû, notamment, à la nature asynchrone de la communication (messages de "correction", phrases communiquées "par morceaux", question-réponse en décalage, etc.). De même, il est nécessaire de prendre en compte la présence d'éléments non-linguistiques, comme des hyperliens, des marqueurs de balisage (HTML ou autre) ou encore des pictogrammes de nature diverses (émoticônes, *smileys* ou *emojis*). De ce fait, il est nécessaire d'effectuer plusieurs étapes de pré-traitement avant de pouvoir les utiliser. En s'inspirant de l'anonymisation déjà effectuée (portant principalement sur les données à caractère personnel du client, par exemple, ses numéros de téléphone sont remplacés par des jetons `_NUMTEL_`), on remplace ces différents éléments par des marqueurs simplifiés (`_HTML_` pour des éléments de balisage HTML, etc.) qui permettent aux modèles d'*embedding* de mots et de tours de parole de les prendre en compte d'une manière similaire à des éléments de ponctuation, sans se soucier de leur sémantique particulière. Dans un même temps, on encode l'identité de l'auteur par un marqueur au début de son message (`___#TC#___` et `___#CLIENT#___` pour le télé-conseiller et le client, respectivement), et qui seront prises en compte dans la représentation du tour de parole.

2. Les données étant mises à disposition des participants du projet ANR Datcha.

| | |
|---|---|
| Nombre de conversations dans le corpus complet | 432 768 |
| Nombre de conversations annotées (état résolution) | 2775 |
| Nombre de tours de paroles dans le corpus complet | 15 682 118 |
| Taille du vocabulaire pour *word embeddings* | 120 391 |
| Dimensions des *word embeddings* | 100 |
| Dimensions cachées de l'encodeur *Skip-Thoughts* | 1024 |
| Structure de l'encodeur *Skip-Thoughts* | biLSTM, 1 couche, 0.5% *dropout* |
| Structure des décodeurs *Skip-Thoughts* (entraînement) | 2 biGRU (suivant, précédent), 1 couche |
| Paramètres de l'entraînement | 10 *epochs*, taille des *batchs* = 32, optimiseur Adam |

TABLE 1 – Description des données et des paramètres du modèle présenté.

# 4 Expériences et résultats

Pour tester notre méthode, au vu du fait que le corpus n'est pas annoté d'une manière directement pertinente pour une mesure de similarité, on opte pour une évaluation indirecte via une tâche externe et des annotations détachées. On fait donc l'hypothèse que la mesure de similarité ainsi produite par notre approche peut potentiellement être corrélée avec l'appartenance à une catégorie d'état de résolution des conversations, qui ont été annotées sur une petite partie du corpus (2775 conversations, contre 432768 pour le corpus d'entraînement total) par l'opérateur de téléphonie. On effectue donc une tâche d'évaluation externe de classification portant sur ces labels, en utilisant un classifieur $k$-plus proches voisins (avec $k = 5$ ici) auquel on fournit une matrice de distance produite par différentes variantes de notre méthode, et sur lesquelles on effectue une validation croisée sur *10-fold* (le jeu de données étant relativement petit). On dispose de 14 classes d'états de résolution annotées, la classe majoritaire représentant $26.19\%$ des données. Les cinq variantes évaluées et comparées ici (table 2) utilisent toutes les mêmes *embeddings* de messages produit par le modèle *FastText* comme base, les différences portant sur la production des représentations des messages, la phase d'alignement et la mesure de distance inter-conversations : la variante *baseline* utilise la moyenne des vecteurs d'*embedding* de messages produits par le modèle *Skip-Thoughts* comme représentation vectorielle des conversations, puis effectue une simple mesure de similarité cosinus comme entrée du classifieur. Les variantes *SIF (Smooth Inverse Frequency)* implémentent la méthode décrite dans (Arora *et al.*, 2017) pour construire des représentations de messages (une version moyenne comme la *baseline*, l'autre avec l'approche *EMD*). Les deux dernières variantes implémentent l'approche basée sur le problème *EMD* présentée dans cet article (avec vecteurs de messages *Skip-Thoughts*), l'une employant la norme $L2$ des vecteurs d'*embedding* de messages comme matrice de coûts pour le solveur, l'autre la similarité cosinus. La variante *baseline* peut paraître excessivement brutale, mais il semblerait qu'au moins au niveau des phrases, l'approche *CBOW (Continuous Bag Of Words)* consistant à effectuer une moyenne des représentations vectorielles capture déjà une partie suprenante de l'informations sémantiques dans les phrases, comme il est montré dans (Adi *et al.*, 2017; Shen *et al.*, 2018).

| **Modèle** | **Exactitude (*accuracy*)** |
|---|---|
| Skip-Thoughts + cos + EMD | **52.53%** ($\sigma \approx 3.36\%$) |
| Skip-Thoughts + L2 + EMD | 51.20% ($\sigma \approx 3.14\%$) |
| SIF + cos + EMD | 48.78% ($\sigma \approx 2.66\%$) |
| SIF + cos + moyenne | 44.73% ($\sigma \approx 1.71\%$) |
| Skip-Thoughts + cos + moyenne (*baseline*) | 43.41% ($\sigma \approx 2.62\%$) |

TABLE 2 – Résultats de la tâche de classification externe sur différentes variantes (*accuracy* moyenne et écart-type sur la validation croisée.)

On observe que la variante *EMD* avec similarité cosinus affiche les meilleures performances, avec une amélioration moyenne de $+9\%$ sur le modèle *baseline*. La similarité cosinus est en général l'opération de prédilection pour comparer des vecteurs d'*embedding* car elle n'est pas sensible à leurs normes, qui, due à la manière dont ceux-ci sont construits, traduit en général une notion analogue à la fréquence d'apparition de l'élément associé dans le corpus, mesure qui n'est en général pas vraiment pertinente quand on souhaite effectuer des comparaisons sémantiques. Ceci peut expliquer les petites différences de performances entre la variante utilisant la norme $L2$ et celle utilisant la similarité cosinus. La distance de Wasserstein, quant à elle, est beaucoup plus sensible aux détails structurels de la conversation : en effet, avec le modèle *baseline* "sac de mots", on peut imaginer que deux conversations structurellement différentes puissent avoir des vecteurs moyens proches, tandis qu'il serait très improbable de trouver deux conversations différentes ayant des distributions de messages très similaires dans l'espace d'*embedding*.

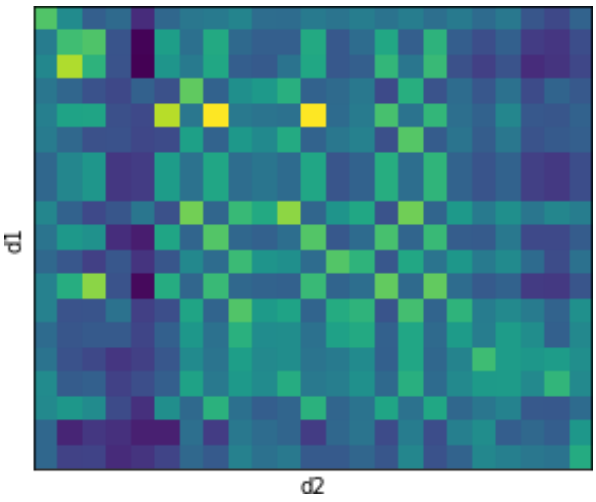
| **I1** | **D1** | **I2** | **D2** |
|---|---|---|---|
| 1 | __#TC#__ Bonjour Mme _CLIENT_ | 1 | __#TC#__ _HTML_ Bonsoir , je me prénomme _TC1_ et je vais traiter votre demande . En quoi puis -... |
| 2 | __#CLIENT#__ Je voudrais ma facture pour le _NUMTEL_ | 3 | __#CLIENT#__ et je veu remplir le formulaire de rembourssement |
| 3 | __#CLIENT#__ comment faire j' ai oublié mon mot de passe | 2 | __#CLIENT#__ bonsoir j ai oublier mon mot de passe svp |
| 4 | __#TC#__ Souhaitez - vous récupérer le mot de passe de votre adresse mail afin de consult... | 7 | __#TC#__ _HTML_ Votre demande consiste t - elle a récupérer le mot de passe de votre adresse de ... |
| 5 | __#CLIENT#__ oui | 8 | __#CLIENT#__ oui |

FIGURE 2 – Extrait d'un alignement optimal produit par notre approche, entre tours de parole de deux conversations (`d1` et `d2`). **I1** et **I2** correspondent respectivement aux indices originaux des interventions dans les conversations (la conversation `d2` étant aligné sur `d1`).

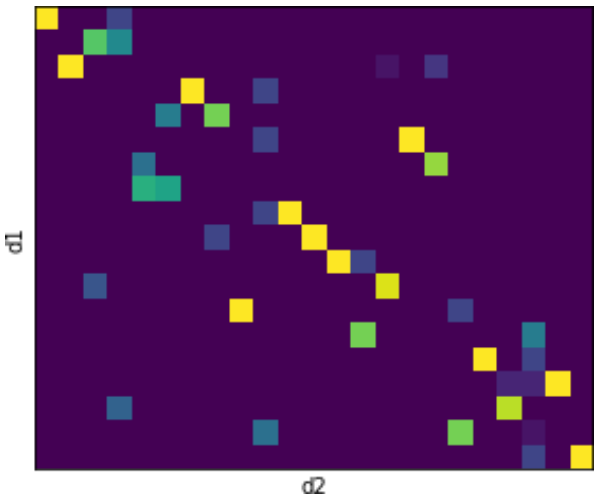
L'approche avec *EMD* est également beaucoup plus interprétable, grâce aux alignements de conversations construits. L'interprétabilité étant un des défis majeurs de la recherche en intelligence artificielle aujourd'hui, la possibilité d'extraire une forme d'explication humainement interprétable, ainsi que les différents objets mathématiques utilisés pour arriver à cette décision, sont des avantages importants de cette approche : dans la figure 2 on peut observer un alignement optimal d'une conversation **D1** avec une conversation similaire **D2**. On peut par exemple voir ici une instance de sémantique structurelle : dans le message de **D1** d'indice 2, le client énonce sa demande principale (récupérer sa facture), suivie en indice 3 du problème instrumental à sa résolution (perte de son mot de passe), tandis que dans **D2**, ces actes de dialogue apparaissent dans l'ordre opposé. Malgré la différence de demande (dans **D2**, le client souhaite remplir un formulaire de remboursement) et l'ordre inversé du déroulement de la conversation, l'alignement produit nous permet qualitativement de juger en quoi celles-ci sont similaires. Une autre manière d'obtenir des interprétations est de directement observer les matrices manipulées par la méthode (figure 3), en particulier la matrice d'alignement solution du problème *EMD* (3b), qui nous permet de rapidement identifier les messages fortement similaires (avec flux important) et les segments de conversation communs mais potentiellement transposés (structures en "diagonales").

L'évaluation présentée ici est bien sûr préliminaire, dans la mesure où le modèle est très simple par rapport à la tâche considérée, et celle-ci ne correspond pas de toutes façons aux applications visées par la similarité. Ces dernières nécessitent une mise en place plus complexe vis à vis de la plate-forme

de conseil, qui n'a pas encore été mise en oeuvre dans le projet global.




(a) Matrice de similarité (en cosinus) entre les représentations sémantiques latentes des messages des deux conversations (jaune : similarité élevée ; bleu : similarité faible).




(b) Matrice solution du problème *EMD* associé, correspondant au meilleur alignement possible entre les deux conversations (jaune : flux important ; bleu : flux faible).

FIGURE 3 – Exemples de matrices associées aux dialogues `d1` et `d2`.

# 5 Perspectives et conclusion

Nous avons proposé ici un modèle qui permet de définir une similarité entre dialogues en prenant en compte le contenu sémantique tout en respectant la structure du dialogue en tours de parole. Il est évident que l'organisation dialogique peut être modélisé de nombreuses façons plus précises, en prenant en compte le type des actes de dialogue (Bunt *et al.*, 2010) ou l'organisation des liens entre tours en fonction des besoins de la communication avec des relations dialogiques, comme dans (Asher *et al.*, 2016). Sans aller jusqu'à ce dernier niveau complexe à prédire même avec des données annotées, ajouter l'information du type d'acte de dialogue (assertion, question, ...) est une suite naturelle : ce niveau est l'objet de nombreux travaux avec des performances assez élevées, que ce soit sur l'anglais (Kumar *et al.*, 2018) ou le français (Perrotin *et al.*, 2018). On pourrait alors observer si l'ajout de cette information à chaque énoncé permet d'avoir de meilleurs alignements et une meilleure correspondance entre des types de dialogue. À l'inverse, il pourrait être intéressant de voir l'apport de l'appariement de dialogues dans l'apprentissage d'une séquence d'actes de dialogues. De plus, si en théorie le modèle *Skip-Thoughts* utilisé ici permet de capturer le contexte immédiat des messages, comprenant donc en partie les spécificités liées à la nature des conversations par *chat*, une amélioration possible à explorer serait l'augmentation de la taille de la fenêtre de contexte à l'entraînement (ici de taille 1 seulement), ou bien, l'utilisation de modèles avec prédictions contextualisées, comme *ELMo* (Peters *et al.*, 2018) ou *BERT* (Devlin *et al.*, 2018). Au-delà du dialogue par *chat*, d'autres formes de communication impliquent de modéliser des interactions textuelles, y compris dans un contexte orienté-tâche, et il serait intéressant de généraliser l'approche à la modélisation de dialogue sur des forums, où (Wang *et al.*, 2012) a montré l'intérêt d'une analyse (supervisée) de la structure.

# Remerciements

Ce travail a été partiellement financé par l'Agence Nationale pour la Recherche dans le cadre du projet ANR-15-CE23-0003 (DATCHA).

# Références

(2017). *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

ADI Y., KERMANY E., BELINKOV Y., LAVI O. & GOLDBERG Y. (2017). Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In (DBL, 2017).

ARORA S., LIANG Y. & MA T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In (DBL, 2017).

ASHER N., HUNTER J., MOREY M., BENAMARA F. & AFANTENOS S. D. (2016). Discourse structure and dialogue acts in multiparty dialogue : the STAC corpus. In *LREC* : European Language Resources Association (ELRA).

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

BUNT H., ALEXANDERSSON J., CARLETTA J., CHOE J., FANG A. C., HASIDA K., LEE K., PETUKHOVA V., POPESCU-BELIS A., ROMARY L., SORIA C. & TRAUM D. R. (2010). Towards an ISO standard for dialogue act annotation. In *LREC* : European Language Resources Association.

CER D., DIAB M., AGIRRE E., LOPEZ-GAZPIO I. & SPECIA L. (2017). Semeval-2017 task 1 : Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, p. 1–14 : Association for Computational Linguistics.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

KIROS R., ZHU Y., SALAKHUTDINOV R., ZEMEL R. S., URTASUN R., TORRALBA A. & FIDLER S. (2015). Skip-thought vectors. In *Advances in Neural Information Processing Systems 28 : Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, p. 3294–3302.

KUMAR H., AGARWAL A., DASGUPTA R. & JOSHI S. (2018). Dialogue act sequence labeling using hierarchical encoder with CRF. In *AAAI*, p. 3440–3447 : AAAI Press.

KUSNER M. J., SUN Y., KOLKIN N. I. & WEINBERGER K. Q. (2015). From word embeddings to document distances. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, p. 957–966 : JMLR.org.

LE Q. V. & MIKOLOV T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, p. 1188–1196.

NAKOV P., HOOGEVEEN D., MÀRQUEZ L., MOSCHITTI A., MUBARAK H., BALDWIN T. & VERSPOOR K. (2017). Semeval-2017 task 3 : Community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, p. 27–48 : Association for Computational Linguistics.

PERROTIN R., NASR A. & AUGUSTE J. (2018). Dialog Acts Annotations for Online Chats. In *25e conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Rennes, France.

PETERS M., NEUMANN M., IYYER M., GARDNER M., CLARK C., LEE K. & ZETTLEMOYER L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 2227–2237, New Orleans, Louisiana : Association for Computational Linguistics.

SHEN D., WANG G., WANG W., RENQIANG MIN M., SU Q., ZHANG Y., LI C., HENAO R. & CARIN L. (2018). Baseline needs more love : On simple word-embedding-based models and associated pooling mechanisms. In *ACL*.

SUTSKEVER I., VINYALS O. & LE Q. V. (2014). Sequence to sequence learning with neural networks. In *NIPS*, p. 3104–3112.

WAN X. (2007). A novel document similarity measure based on earth mover's distance. *Information Sciences*, **177**(18), 3718–3730.

WANG L., KIM S. N. & BALDWIN T. (2012). The utility of discourse structure in identifying resolved threads in technical user forums. In *COLING*, p. 2739–2756 : Indian Institute of Technology Bombay.

YANG Y., YUAN S., CER D., KONG S.-Y., CONSTANT N., PILAR P., GE H., SUNG Y.-H., STROPE B. & KURZWEIL R. (2018). Learning semantic textual similarity from conversations. In *Proceedings of The Third Workshop on Representation Learning for NLP*, p. 164–174 : Association for Computational Linguistics.

# Résolution des coréférences neuronale : une approche basée sur les têtes

Quentin Gliosca[1,2] * Pascal Amsili[3]
[1] École Polytechnique Fédérale de Lausanne
[2] Haute École d'Ingénierie et de Gestion du Canton de Vaud
[3] Laboratoire de Linguistique Formelle (Université Paris Diderot & CNRS)
[1] quentin.gliosca@gmail.com
[3] amsili@linguist.univ-paris-diderot.fr

RÉSUMÉ

L'avènement des approches neuronales de bout en bout a entraîné une rupture dans la façon dont était jusqu'à présent envisagée et implémentée la tâche de résolution des coréférences. Nous pensons que cette rupture impose de remettre en question la conception des mentions en termes de syntagmes maximaux, au moins pour certaines applications dont nous donnons deux exemples. Dans cette perspective, nous proposons une nouvelle formulation de la tâche, basée sur les têtes, accompagnée d'une adaptation du modèle de Lee *et al.* (2017) qui l'implémente.

ABSTRACT

**Neural coreference resolution : a head-based approach**

End-to-end neural approaches for coreference resolution broke away from the traditional implementation and conception of this task. We argue that, at least for some applications, this breakaway forces us to reconsider the definition of mentions in terms of maximal syntagms : we consider the examples of machine translation and summarization. In this perspective, we propose a new formulation of the task, based on mention heads, and adapt the model of Lee *et al.* (2017) to that end.

MOTS-CLÉS : résolution des coréférences, réseaux de neurones, modèles de bout en bout.

KEYWORDS: coreference resolution, neural networks, end-to-end models.

## 1 Introduction

La tâche de résolution des coréférences consiste à partitionner des mentions de référents de discours en chaînes de coréférence. Cette définition abstraite ne spécifie pas la nature exacte des mentions de référents de discours, qui demeure à ce jour mystérieuse, et pose la question de la forme linguistique sous laquelle les mentions apparaissent et grâce à laquelle elles peuvent être localisées dans un texte.

C'est une représentation syntaxique des mentions qui a été conventionnellement choisie en TAL : on prend pour définition d'une mention un empan de texte correspondant à un syntagme maximal, c'est-à-dire à la projection maximale d'un des mots de la phrase. Ainsi, la tâche de résolution des coréférences telle qu'elle est pratiquée correspond d'une part à une analyse syntaxique et d'autre part à la résolution des coréférences à proprement parler.

∗. Cet article fait suite à un travail réalisé à l'Université Paris Diderot.

Ce choix n'a pas été sans conséquence sur les architectures des systèmes de résolution des coréférences proposées. Il nous apparaît que l'émergence des architectures neuronales de bout en bout — comme celle de Lee *et al.* (2017) — invite aujourd'hui à le rediscuter.

Après avoir évoqué les deux principales architectures de système de résolution des coréférences neuronaux (§ 2), nous nous intéresserons à la section 3 à la nature des attentes que peuvent avoir les applications aval par rapport à un système de coréférences (en ne mentionnant brièvement, faute de place, que quelques exemples) avant de proposer une approche de bout en bout basée sur les têtes, dont les premiers résultats que nous rapportons suggèrent qu'elle peut être fructueuse (§ 4).

# 2 Approches neuronales existantes

## 2.1 Approche en deux étapes

À la suite de Soon *et al.* (2001), de nombreux modèles statistiques de résolution des coréférences ont été proposés. La majorité d'entre eux utilisait des classeurs ou ordonneurs linéaires pour déterminer des liens de coréférence entre mentions, voire entre une mention et une chaîne de coréférence partielle dans le cas des modèles basés sur les entités.

Ces modèles à l'expressivité limitée impliquaient d'extraire indépendamment les descripteurs pertinents pour résoudre les liens de coréférence. Aussi, les architectures proposées procédaient typiquement à une analyse syntaxique préalable à la résolution des coréférences pour extraire à la fois les mentions et des descripteurs pertinents, au premier rang desquels leurs têtes.

Cette architecture en deux étapes, autour de laquelle a longtemps été conçue la résolution des coréférences, permettait de disposer à la fin du processus à la fois des liens de coréférence entre mentions représentées par des empans maximaux, et des représentations linguistiques riches de ces mentions sous forme d'arbres syntaxiques.

Suite aux premiers succès de l'apprentissage de représentations neuronales en TAL, Wiseman *et al.* (2015) ont proposé le premier modèle neuronal de résolution des coréférences, qui était donc capable d'apprendre à calculer des représentations vectorielles des mentions pertinentes pour la tâche, avant de procéder à la traditionnelle étape de résolution des coréférences par ordonnement linéaire.

Ces travaux ont été suivis de plusieurs autres propositions de modèles neuronaux (Clark & Manning, 2015, 2016; Wiseman *et al.*, 2016), qui ont permis de très rapidement faire progresser l'état de l'art de plusieurs points.

## 2.2 Approche de bout en bout

Le modèle introduit par Lee *et al.* (2017) se démarque des systèmes antérieurs en abandonnant l'architecture en deux étapes pour pleinement tirer parti des capacités des réseaux de neurones. Grâce à leur expressivité, ceux-ci sont en effet capables d'apprendre implicitement les descripteurs pertinents, y compris ce qui différencie une mention d'un empan de texte quelconque, rendant de ce fait l'analyse syntaxique implicite.

Concrètement, le modèle considère tous les empans à l'intérieur de chaque phrase comme membre

potentiel d'une relation de coréference. Cela revient à joindre la tâche de détection des mentions et la tâche de résolution des coréférences. Cette approche a permis un gain de performance de 1,5 points CoNLL que les auteurs attribuent à la réduction de la propagation d'erreur typique d'une architecture en pipeline.

Cette approche a cependant comme défaut de produire un résultat beaucoup moins riche que l'architecture en deux étapes : des chaînes de coréférence d'empans maximaux plats, sans aucune information supplémentaire.

Il semble maintenant utile de prendre un peu de recul par rapport à la tâche normalisée : en effet, les performances du modèle de Lee *et al.* (2017) pourraient perdre de leur intérêt si sa sortie se révélait insuffisante pour les applications aval. Nous considérerons brièvement quelques exemples dans la section suivante.

# 3 Interlude : quelques exemples d'application

De nombreuses tâches qui sont ou pourraient être utilisatrices des résultats d'un résolveur de coréférences ne sont pas intéressées par les bornes des mentions telles que définies dans la tâche standard. On peut en effet distinguer deux principaux cas d'utilisation du résultat d'un résolveur de coréférences :

— on souhaite savoir quelles chaînes de coréférence sont présentes dans chacune des phrases, ou tout au plus dans quelle phrase est situé l'antécédent de telle anaphore ;
— on souhaite connaître les empans des mentions, mais aussi leur structure linguistique interne, par exemple pour transformer le texte.

Dans le premier cas, aucune localisation précise des mentions n'est requise et elles peuvent être identifiées de n'importe quelle manière. Dans le second, les délimitations en termes de syntagmes maximaux s'avèrent souvent insuffisantes, l'information requise pouvant aller jusqu'aux arbres syntaxiques complets des mentions.

Sans prétendre si brièvement couvrir de façon représentative l'ensemble des applications utilisatrices de résolution de coréférences, nous allons maintenant décrire une approche de résumé automatique, et une de traduction automatique qui permettent d'illustrer ces considérations.

## 3.1 Résumé automatique

Le résumé automatique extractif construit le résumé d'un document par sélection de certaines de ses phrases jugées particulièrement informatives, en général par une fonction de score. Durrett *et al.* (2016) remarquent que, sans précautions particulières, 60 % des phrases choisies contiennent des pronoms orphelins. Pour remédier à ce problème, ils proposent d'ajouter des contraintes d'anaphoricité dans leur système par deux approches complémentaires, toutes deux basées sur un système de résolution des coréférences.

La première méthode s'applique lorsque, pour un pronom donné, une chaîne de coréférence est prédite par le système avec une forte probabilité. Si une mention de l'entité n'a pas encore été incluse dans le résumé, le pronom est remplacé par la première mention de sa chaîne de coréférence. Dans le cas

contraire (deuxième méthode), le système force l'inclusion dans le résumé de contenu supplémentaire, de façon à garantir la clarté de la référence du pronom.

Dans ce dernier cas, les syntagmes maximaux ne jouent aucun rôle et il suffirait amplement, par exemple, d'identifier les différentes mentions par leurs têtes. En ce qui concerne le premier cas, une note de bas de page précise que la première mention de la chaîne de coréférence du pronom n'est en fait pas toujours utilisée telle quelle.

En effet, si la tête de la mention est un nom propre, le pronom n'est remplacé que par la partie de la mention qui correspond au nom propre, plutôt que par le syntagme nominal entier. Mais même dans le cas d'un syntagme nominal standard, le remplacement aveugle d'un pronom par un syntagme maximal pourrait conduire à des incohérences. Les syntagmes maximaux ne sont donc pas, dans ce cas-là non plus, les délimitations les plus pertinentes.

De plus, du fait de la grande hétérogénéité des liens qui unissent deux mentions d'une même chaîne de coréférence, de nombreuses précautions doivent être prises avant de procéder à un remplacement. Durrett *et al.* (2016) précisent qu'ils veillent par exemple à remplacer les adjectifs possessifs par un syntagme possessif.

Cet exemple illustre d'une part, que les besoins en matière de résolution des coréférences sont très variés et que des empans de texte bien délimités ne sont pas toujours nécessaires ; d'autre part, que lorsque des délimitations sont requises, il ne s'agit non seulement pas nécessairement des syntagmes maximaux, mais qu'en plus, beaucoup d'informations sur les mentions sont nécessaires pour pouvoir exploiter efficacement les chaînes de coréférence.

Typiquement, l'analyse syntaxique traditionnellement effectuée en amont de la résolution des coréférences permettrait ici d'identifier les noms propres [1] et de supprimer les propositions relatives des mentions. La gestion des possessifs pourrait quant à elle éventuellement se contenter de quelques heuristiques sur les mots compte tenu de la grande simplicité de l'anglais en la matière.

## 3.2 Traduction automatique

À la suite de Le Nagard & Koehn (2010), l'utilisation de la résolution des coréférences en traduction automatique a été abondamment décrite. Les principaux modèles de traduction automatique traduisent les phrases du document source une par une indépendamment les unes des autres, ce qui n'est pas sans conséquence sur la cohérence de la traduction.

À titre d'illustration, reprenons l'exemple de Le Nagard & Koehn (2010). *Google Translate* traduisait à l'époque le discours (1a) par (1b). Le pronom *it* est ici mal traduit en l'absence d'information sur le genre de son antécédent.

(1) a. The window is open. It is black.
b. La fenêtre est ouverte. Il est noir.

Pour éviter ce genre d'incohérences qui, dans des cas moins triviaux, nuisent grandement à la compréhensibilité du texte traduit, Le Nagard & Koehn (2010) proposent de prétraiter les documents en amont du modèle de traduction. Un système de résolution des coréférences est appliqué sur le

1. Dans le très utilisé jeu d'étiquettes morphosyntaxiques du Penn Treebank (Marcus *et al.*, 1993), les noms propres sont clairement marqués par les étiquettes *NNP* et *NNPS*.

document source pour identifier les antécédents des pronoms qui sont remplacés par des pronoms factices porteurs d'information sur le genre que doit avoir leur traduction.

Dans ce cas d'utilisation de la résolution des coréférences, c'est la tête des mentions qui est utile puisque c'est elle qui est porteuse du genre du référent. Un syntagme maximal sans analyse syntaxique pour identifier la tête de la mention s'avérerait de fait inexploitable.

# 4 Approche de bout en bout basée sur les têtes

## 4.1 Principe général

D'après ce qui précède, il est parfois souhaitable de disposer des têtes des mentions plutôt que d'empans maximaux. Dans l'approche traditionnelle de la résolution des coréférences par étapes successives d'analyse, cela ne pose aucune difficulté : les mentions sont extraites en amont avec leurs arbres syntaxiques complets dont on peut facilement extraire les têtes. En revanche, dans l'approche de bout en bout basée sur les empans proposée par Lee *et al.* (2017), seuls des empans maximaux plats sont extraits du texte.

Une solution naïve consisterait à appliquer une analyse syntaxique sur le texte après la résolution des coréférences pour extraire les têtes des mentions, mais cela reviendrait à faire deux fois l'analyse syntaxique : une fois implicitement et sous les fortes contraintes de complexité propres à la résolution des coréférences, pour trouver les empans maximaux, et une seconde fois explicitement pour trouver leurs têtes.

Nous avons cependant expérimenté une approche différente : travailler directement sur les têtes, c'est-à-dire choisir les têtes sémantico-syntaxiques plutôt que les syntagmes maximaux pour représenter les mentions. Notons d'ailleurs que c'est le choix qui est à peu près toujours fait dans le cas particulier de la résolution des anaphores événementielles.

Il semble bien sûr impossible de faire une résolution des coréférences de qualité en se basant uniquement sur des mots, mais les architectures neuronales permettent aujourd'hui de contextualiser les mots très efficacement. Par exemple, le modèle de Lee *et al.* (2017) peut être très simplement adapté pour produire des chaînes de coréférence de têtes puisque chaque token est contextualisé par un réseau récurrent bidirectionnel (Schuster & Paliwal, 1997).

La question de l'évaluation d'un système basé sur les têtes est plus simple puisqu'il n'est plus question de décider arbitrairement si des bornes de mentions partiellement correctes, voire approximatives, doivent être sanctionnées plus ou moins sévèrement. [2] Les métriques usuelles d'évaluation de la résolution des coréférences peuvent être utilisées de la même manière que sur des empans maximaux, avec l'avantage que seule est évaluée la capacité du modèle à détecter la référentialité et à résoudre les relations de coréférences entre les mentions, indépendamment de toute considération syntaxique.

2. Les erreurs dans les bornes des mentions prédites ont été traitées avec une sévérité plus ou moins grande selon l'époque et le corpus utilisé. Lors des campagnes d'évaluation MUC, une prédiction était considérée comme correcte dès lors que la mention prédite contenait la tête annotée et qu'elle ne dépassait pas du syntagme maximal. La plus récente campagne SemEval 2010 s'est montrée un peu plus sévère : 1 point était accordé si la mention prédite coïncidait parfaitement avec celle annotée, mais seulement 0.5 dans le cas où elle contenait sa tête et ne s'étendait pas au-delà des bornes de référence. Enfin, les campagnes d'évaluation CoNLL 2011 et CoNLL 2012 ont introduit la règle en vigueur aujourd'hui : une mention prédite n'est considérée comme correcte que si ses bornes coïncident exactement avec celles de référence.

## 4.2 Modèle

Pour construire un modèle de résolution des coréférences de bout en bout basé sur les têtes, il suffit de considérer chaque paire de mots comme potentiellement coréférente, plutôt que chaque paire d'empans de phrase. La complexité intrinsèquement quartique de l'approche de Lee *et al.* (2017) qui avait nécessité de nombreuses simplifications pour se ramener à une complexité linéaire (par exemple limiter arbitrairement la longueur des mentions), est de fait ramenée à une complexité intrinsèquement quadratique. La seule hypothèse simplificatrice qui reste nécessaire pour atteindre une complexité linéaire est de limiter la distance à laquelle peuvent se trouver deux mentions coréférentes.

Concrètement, seule la représentation des candidats mentions diffère du modèle basé sur les empans. Dans l'approche proposée par Lee *et al.* (2017), le $i$-ème empan est représenté par le vecteur $g_i$, concaténation des vecteurs contextualisés des premier et dernier mots de l'empan, d'un vecteur d'attention à la tête, et d'un vecteur $\phi(i)$ encodant la largeur de l'empan (équation 1).

$$g_i = \left[x^*_{DEBUT(i)}; x^*_{FIN(i)}; \hat{x}_i; \phi(i)\right] \tag{1}$$

Dans notre modèle basé sur les têtes, les candidats mentions sont les mots du texte plutôt que des empans. Aussi, le mot $i$ est simplement représenté par le vecteur contextualisé qui lui correspond, le mécanisme d'attention à la tête étant naturellement abandonné (équation 2).

$$g_i = x^*_i \tag{2}$$

Puisque le vecteur $x^*_i$ est une représentation contextualisée du $i$-ème mot, vraisemblablement riche en informations syntaxiques et sémantiques, notre approche va cependant bien au-delà de l'heuristique de la même tête proposée par Elsner & Charniak (2010).

## 4.3 Expériences

Pour évaluer l'intérêt de notre approche, nous comparons notre modèle directement basé sur les têtes à une approche naïve utilisant le modèle basé sur les empans de Lee *et al.* (2017) puis une extraction des têtes à partir des arbres syntaxiques de référence.

Les deux modèles utilisés sont entraînés et évalués sur le corpus CoNLL 2012 (Pradhan *et al.*, 2012), basé sur Ontonotes (Hovy *et al.*, 2006). Ce corpus ne fournissant pas les têtes des mentions, nous les avons déterminées en utilisant une version légèrement modifiée de l'algorithme de recherche de têtes *ModCollinsHeadFinder* (Collins, 1999) [3] de la boîte à outils *Stanford CoreNLP* (Manning *et al.*, 2014). Les modifications apportées sont les suivantes :

(i) les conjonctions de coordination sont considérées comme les têtes des syntagmes nominaux coordonnés pour s'assurer qu'un nom ne peut pas être à la fois la tête du syntagme coordonné et d'un des conjoints ;

(ii) si aucune règle ne s'applique au constituant, son fils le plus à droite est sélectionné. Cet aménagement est nécessaire pour gérer les étiquettes comme *EMBED* pour lesquelles aucune règle n'existe ;

3. Cet algorithme détermine récursivement la tête d'un constituant à partir des étiquettes syntaxiques de ses enfants.

(iii) dans le rare cas où une mention n'est pas un constituant syntaxique, le mot le plus à gauche de l'empan est arbitrairement considéré comme sa tête.

Les têtes ainsi obtenues avec la syntaxe de référence sont utilisées dans nos expériences pour :

(i) construire les exemples d'apprentissage du modèle basé sur les têtes ;

(ii) produire un corpus d'évaluation annoté en têtes ;

(iii) transformer les empans maximaux produits par le modèle basé sur les empans, dans l'évaluation de l'approche naïve.

Les scores reportés dans le tableau 1 montrent qu'un important gain de performance se dessine en rappel grâce à la prise en compte de toutes les mentions, sans limite de taille imposée aux syntagmes maximaux, et induit des performances globales supérieures de 1,6 points en termes de F1 CoNLL. Notons également que grâce à la baisse de complexité intrinsèque déjà mentionnée, les temps de calcul et l'empreinte mémoire du modèle basé sur les têtes sont diminués approximativement d'un facteur deux comparé au modèle basé sur les empans de Lee *et al.* (2017).

| | MUC | | | $B^3$ | | | $CEAF_e$ | | | CoNLL | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Prec. | Rec. | F1 | Prec. | Rec. | F1 | Prec. | Rec. | F1 | Prec. | Rec. | F1 |
| Notre approche | 78.12 | **77.84** | **77.98** | 68.37 | **67.3** | **67.83** | 63.31 | **64.32** | **63.81** | 69.94 | **69.80** | **69.87** |
| Approche naïve | **79.62** | 74.29 | 76.86 | **70.08** | 62.65 | 66.15 | **63.66** | 59.91 | 61.73 | **71.12** | 65.62 | 68.25 |

TABLE 1 – Scores de coréférences entre têtes sur le corpus de test CoNLL 2012.

# 5 Conclusion

La tâche de résolution des coréférences telle qu'on la connaît aujourd'hui n'est qu'une formulation parmi d'autres du problème linguistique sous-jacent. En fonction des applications utilisatrices, des formulations et modèles alternatifs plus appropriés devraient être envisagés.

Nous proposons par exemple une attrayante approche de la tâche, basée sur les têtes, qui pourrait trouver sa place dans plusieurs applications. Cette formulation alternative permet en outre de se focaliser sur la résolution des coréférences à proprement parler en se débarrassant de considérations syntaxiques pas toujours pertinentes.

Comme preuve de concept, nous montrons que lorsqu'il s'agit d'obtenir les têtes des mentions dans l'approche de bout en bout, il est plus efficace de légèrement adapter le modèle proposé par Lee *et al.* (2017) que de post-traiter par analyse syntaxique les empans maximaux qu'il produit.

# Remerciements

Ce travail a reçu le soutien du *Labex* EFL (*Empirical Foundations of Linguistics*, ANR-10-LABX-0083). Nous remercions Marie Candito, Olga Seminck, Loïc Grobol, Benoît Crabbé pour leurs commentaires critiques et constructifs sur des versions antérieures de cet article, ainsi que les relecteurs de TALN qui ont fourni un retour précieux. Nous remercions enfin Timothee Mickus pour son aide dans la mise en forme de versions précédentes de ce travail.

# Références

CLARK K. & MANNING C. (2015). Entity-Centric Coreference Resolution with Model Stacking. In *Proceedings of the 53rd Annual Meeting of the ACL*.

CLARK K. & MANNING C. (2016). Deep Reinforcement Learning for Mention-Ranking Coreference Models. In *Proceedings of EMNLP 2016*.

COLLINS M. (1999). *Head-Driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania.

DURRETT G., BERG-KIRKPATRICK T. & KLEIN D. (2016). Learning-Based Single-Document Summarization with Compression and Anaphoricity Constraints. In *Proceedings of the 54th Annual Meeting of the ACL*.

ELSNER M. & CHARNIAK E. (2010). The Same-Head Heuristic for Coreference. In *Proceedings of the ACL 2010 Conference*.

HOVY E., MARCUS M., PALMER M., RAMSHAW L. & WEISCHEDEL R. (2006). OntoNotes : The 90% Solution. In *Proceedings of the Human Language Technology Conference of the NAACL*.

LE NAGARD R. & KOEHN P. (2010). Aiding Pronoun Translation with Co-Reference Resolution. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation*.

LEE K., HE L., LEWIS M. & ZETTLEMOYER L. (2017). End-to-end Neural Coreference Resolution. In *Proceedings of EMNLP 2017*.

MANNING C., SURDEANU M., BAUER J., FINKEL J., BETHARD S. & MCCLOSKY D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the ACL*.

MARCUS M. P., MARCINKIEWICZ M. A. & SANTORINI B. (1993). Building a large annotated corpus of English : The Penn Treebank. *Computational linguistics*.

PRADHAN S., MOSCHITTI A., XUE N., URYUPINA O. & ZHANG Y. (2012). CoNLL-2012 shared task : Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*.

SCHUSTER M. & PALIWAL K. (1997). Bidirectional Recurrent Neural Networks. *Transactions on Signal Processing*.

SOON W. M., NG H. T. & LIM D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*.

WISEMAN S., RUSH A. & SHIEBER S. (2016). Learning Global Features for Coreference Resolution. In *Proceedings of the 2016 Conference of the NACACL*.

WISEMAN S., RUSH A., SHIEBER S. & WESTON J. (2015). Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of the 53rd Annual Meeting of the ACL and the 7th IJCNLP*.

# Réutilisation de textes dans les manuscrits anciens

Amir Hazem[1] Béatrice Daille [1] Dominique Stutzmann [2] Jacob Currie [2]
Christine Jacquin [1]
(1) LS2N, 2 Chemin de la Houssinière, 44322 Nantes
(2) IRHT, 40 Avenue d'Iéna, 75116 Paris
amir.hazem@ls2n.fr, beatrice.daille@ls2n.fr,
dominique.stutzmann@irht.cnrs.fr, jacob.currie@irht.cnrs.fr,
christine.jacquin@ls2n.fr

RÉSUMÉ

Nous nous intéressons dans cet article à la problématique de réutilisation de textes dans les livres liturgiques du Moyen Âge. Plus particulièrement, nous étudions les variations textuelles de la prière *Obsecro Te* souvent présente dans les livres d'heures. L'observation manuelle de 772 copies de l'*Obsecro Te* a montré l'existence de plus de 21 000 variantes textuelles. Dans le but de pouvoir les extraire automatiquement et les catégoriser, nous proposons dans un premier temps une classification lexico-sémantique au niveau n-grammes de mots pour ensuite rendre compte des performances de plusieurs approches état-de-l'art d'appariement automatique de variantes textuelles de l'*Obsecro Te*.

ABSTRACT

**Text Reuse in Ancient Manuscripts**

We address in this paper the issue of text reuse in liturgical books of the middle age. More specifically, we study variant readings of the *Obsecro Te* prayer, part of the devotional Books of Hours. The manual observation of 772 copies of *Obsecro Te* has shown more than 21,000 textual variants. In order to automatically extract and categorize them, we first introduce a semantico-synformic classification at the ngram level, then, we contrast several unsupervised state-of-the-art approaches for the automatic aquisition of *Obsecro Te* variants.

MOTS-CLÉS : Obsecro Te, Livres d'heures, Réutilisation de textes, Variantes textuelles.

KEYWORDS: Obsecro Te, Books of hours, Text reuse, Textual variants.

# 1 Introduction

La religion chrétienne utilise plusieurs types de livres liturgiques. Empruntant leurs principaux éléments à l'un d'eux, le bréviaire, les livres d'heures sont un recueil de prières à l'usage des fidèles (Leroquais, 1927). Souvent richement enluminés, et répandus dès le 13e siècle en France, au sud des Pays-Bas, en Angleterre et plus tard en Italie et en Espagne, ils constituent une part importante de l'ensemble des manuscrits médiévaux préservés et sont une source d'information sur la vie et la chrétienté au Moyen Âge. Ils reproduisent le contenu de livres réservés aux prêtres et au clergé et permettent aux laïques de prier, comme ceux-ci, selon les heures canoniales. Les livres d'heures ont un noyau en latin et des additions en langues vernaculaires (souvent en français) et font partie des textes les plus lus au Moyen Âge. Malgré leur succès à l'époque, il s'avère aujourd'hui que leur contenu

textuel reste très peu étudié. De plus, il existe très peu de livres d'heures transcrits et annotés. L'une des rares ressources sur le texte des livres d'heures est la base *Beyond Use*, qui contient, en particulier, une section sur l'*Obsecro Te* (Plummer & Clark, 2015). Cette prière à la Vierge a été transcrite et annotée manuellement à partir de plus de 772 livres d'heures [1]. Ainsi, plus de 21 000 variantes textuelles ont été enregistrées. Les variantes sont le résultat d'une opération d'addition, suppression ou substitution au niveau du mot. Une même opération regroupe des opérations linguistiques diverses. Ainsi une opération de substitution peut faire référence, entre autres, à des variantes flexionnelles (*crucem / cruce*), des variantes paradigmatiques obtenues par substitution synonymique (*gratie / indulgencie*). Deux opérations de substitution consécutives peuvent caractériser des variantes de permutation (*opera misericordia / misericordia opera*).
Nous abordons dans cet article la tâche d'extraction automatique de variantes textuelles dans les livres d'heures et plus particulièrement en utilisant *l'Obsecro Te* comme ressource d'évaluation. Ce travail qui constitue une première amorce, a pour but à terme, d'étudier le contenu textuel des livres d'heures afin de découvrir leurs différents usages et de déceler des similarités sur différentes granularités. Ces similarités pourraient servir par exemple à détecter des corrélations structurelles, géographiques et terminologiques entre livres d'heures provenant de différentes régions, d'un même pays ou de pays différents dans cette Europe médiévale. Nous examinons ces différences en proposant, dans un premier temps, une classification lexico-sémantique des variantes au niveau n-grammes de mots, pour ensuite rendre compte des performances de plusieurs approches état-de-l'art d'appariement automatique de variantes textuelles de l'*Obsecro Te*.

# 2 État de l'art

Un intérêt grandissant pour le traitement automatique du contenu textuel de manuscrits anciens commence à émerger avec un but majeur, qui est celui de pouvoir associer à la fois des analyses historiques et littéraires sur les réseaux textuels (Léonelli, 1985; Stutzmann, 2015; Dondi, 2016). La détection de variantes textuelles constitue un premier pas dans cette direction et plusieurs approches état-de-l'art dédiées à l'alignement et au plagiat par exemple, peuvent être envisagées. La plupart des approches traitant le plagiat ont été proposées et évaluées lors des campagnes PAN [2] de 2009 à 2015 (Belyy *et al.*, 2018). Parmi les approches les plus efficaces, nous pouvons citer celles à base de modèles par plongements de mots (Brlek *et al.*, 2016), celles utilisant les algorithmes génétiques (Kanhirangat & Gupta, 2016; Sanchez-Perez *et al.*, 2018) ou encore, celles à base de modèles thématiques (Le *et al.*, 2016). D'autres approches à base de réseaux de neurones profonds avec des architectures complexes peuvent aussi être envisagées, par exemple les réseaux à convolutions (CNN) (He *et al.*, 2015). Dans ce présent travail, ce type de modèles est difficilement applicable, d'une part, parce qu'il exploite le phénomène de la paraphrase, ce qui n'est pas ou peu le cas concernant les variantes de textes religieux, et d'autre part, le manque de données d'entraînements à disposition ne permet pas de réaliser un apprentissage efficace. Ainsi, nous abordons principalement des méthodes d'alignement classiques à base de similarité de chaînes de caractères et de mots comme la distance d'édition (Levenshtein, 1966) et l'indice de Jaccard (Jaccard, 1901), les approches distributionnelles (Firth, 1957; Harris, 1971) et les approches par plongements de mots (Brlek *et al.*, 2016; Arora *et al.*, 2017).

1. http://www6.sewanee.edu/beyonduse/
2. https://pan.webis.de

# 3 Variantes textuelles dans *l'Obsecro Te*

Nous présentons une nouvelle catégorisation de variantes inspirée de la similarité lexicale (similar lexical forms ou synforms) introduite par Laufer (1988) et de la typologie de variantes terminologiques proposée par Daille (2017) .

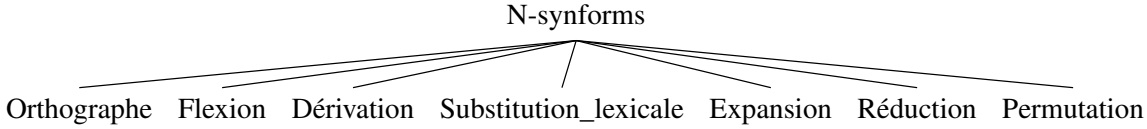
## 3.1 Similarité lexicale (Synforms)

Le concept de formes lexicales similaires (Similar lexical forms ou synforms en Anglais) a été introduit dans le but d'étudier les confusions lexicales des apprenants de l'anglais (Laufer, 1988). Les synforms sont définis au niveau du mot et sont classés selon différentes catégories de similarité comme des variantes : productives de même racine et de suffixes différents (*considerable / considerate, successful / successive*) ; non productives de même racine et de suffixes différents (*credible / credulous, capable / capacious*) ; ayant des consonnes identiques et des voyelles différentes (*base / bias, manual / menial*) ; ayant des phonèmes identiques à l'exception d'une consonne (*price / prize, extend / extent*), etc.

## 3.2 Similarité lexicale au niveau des séquences de mots (N-Synforms)

Nous étendons le concept de formes lexicales similaires (synforms) (Laufer, 1988; Kocic, 2008) au niveau des n-grammes de mots. Cependant, nous n'utilisons pas les 10 catégories présentées dans (Laufer, 1988) puisqu'elles ont été construites sur la base des confusions des apprenants de l'anglais. En revanche, nous conservons les catégories s'appliquant au mot seul (unigramme) communes à celle de Daille (2017) et étendons notre jeu de catégories avec certaines opérations linguistiques caractéristiques des termes complexes et s'appliquant aux n-grammes.

Notre observation à la fois des multiples versions de l'*Obsecro Te* et des annotations de celles-ci à l'aide d'opérations d'édition (Plummer & Clark, 2015) nous conduit à proposer une typologie de variantes motivée linguistiquement et pouvant s'appliquer à des séquences de mots de longueur variable. Notre typologie inclut les variations linguistiques classiques opérant sur le mot (orthographe, flexion et dérivation), la substitution lexicale, ainsi que les opérations spécifiques aux séquences de mots (réduction, expansion et permutation). La figure ci-dessous résume notre typologie :




Nous détaillons maintenant nos catégories de variantes :

**Orthographe** des substitutions de lettres au sein du mot (consonnes ou voyelles) comme *dilecto / delecto* ;

**Flexion** les flexions en cas du latin, comme *crucem / cruce* ;

**Dérivation** toute opération de dérivation morphologique pouvant engendrer ou non un changement de catégorie grammaticale, comme *dilecto (Adj)/ dilectissimo (Adj superlatif)* ;

**Substitution lexicale** toute opération de substitution d'une unité lexicale par une autre. La substitution lexicale permet de générer des variantes en relation de synonymie (*tribuas / concedas*), en relation de quasi-synonymie (*gratie / indulgencie*) mais aussi d'autres variantes sans relation sémantique claire (*tribuas / obtineas*) ;

**Expansion** les opérations linguistiques d'expansion sont la modification et la prédication comme *criminalibus peccatis / criminalibus peccatis vel mortalibus* ;

**Réduction** les opérations linguistiques de réduction sont la réduction lexicale et la réduction anaphorique comme *ostendem michi gloriosam / ostendem michi* ;

**Permutation** la permutation comme *criminalibus peccatis / peccatis criminalibus*.

Bien entendu, comme toute typologie, la nôtre ne prétend pas à l'exhaustivité. Elle pourra être étendue si nécessaire à d'autres opérations linguistiques comme la composition ou la coordination si celles-ci sont rencontrées. Des variantes combinant de multiples opérations, comme des substitutions lexicales associées à des expansions ou des permutations, existent mais elles sont rares.

# 4 Approches

## 4.1 Distance d'édition (Levenshtein)

La distance d'édition (Levenshtein, 1966) mesure la proximité entre deux mots $x$ et $y$ en attribuant un score prenant en compte le nombre d'insertions, de suppressions et de substitutions nécessaires pour transformer $x$ en $y$. Plus le score est élevé, plus le nombre de changements est important et moins les mots sont similaires. Parmi les applications de la distance d'édition, nous retrouvons la détection de plagiat ou la correction orthographique. La formule de la distance d'édition est représentée ci-dessous :

$$D(i,j) = min \begin{cases} D[i-1,j] + SuppCout(i) \\ D[i,j-1] + InsCout(i) \\ D[i-1,j-1] + SubCout(i,j) \end{cases} \quad (1)$$

avec $D(i,j)$ la distance entre les n-grammes $i$ et $j$, et $SuppCout(i)$ la fonction de coût de suppression de $i$, $InsCout(i)$ la fonction de coût d'insertion de $i$ et $SubCout(i,j)$ la fonction de coût de substitution de $i$ par $j$. Pour se ramener à la distance de Levenshtein les trois fonctions de coût sont mises à 1. Nous utilisons cette mesure dans la problématique d'extraction de variantes car certaines variantes latines observées dans *l'Obsecro Te* peuvent être très proches comme *salvatione* avec *salvationis* ou *salvationem*. Dans ce cas, la distance d'édition est très efficace pour détecter ces variantes. Nous obtenons par exemple un score d'édition de 2 entre *salvatione* et *salvationis* (la substitution de la lettre *e* par *i* et l'ajout du *s*) et un score de 1 entre *salvatione* et *salvationem* (1 ajout de la lettre *m*).

## 4.2 Indice de Jaccard

L'indice de Jaccard (Jaccard, 1901) mesure le degré de similarité entre deux ensembles. Ceci est représenté par le nombre d'éléments en commun entre les deux ensembles divisé par la totalité des éléments des deux ensembles. Plus il y a d'éléments en commun plus le score est proche de zéro

et plus les séquences sont similaires. L'un des avantages de l'indice de Jaccard est qu'il ne prend pas en compte la position des éléments dans les deux séquences. Cette mesure est donc efficace pour détecter les variantes de permutation en leur attribuant un score égal à 0. La formule ci-dessous exprime l'indice de Jaccard

$$Jaccard(A, B) = \frac{A \cap B}{A \cup B} \quad (2)$$

où les deux ensembles A et B correspondent à deux n-grammes de mots, avec B une variante candidate. L'intersection comme l'union sont considérées au niveau du caractère.

## 4.3 Adaptation de l'approche distributionnelle

L'approche distributionnelle (par sac de mots) classique consiste à représenter chaque mot par son vecteur de contextes (Firth, 1957; Harris, 1971). Chaque contexte représente les mots qui entourent un mot donné selon une taille de fenêtre. Nous adaptons cette approche aux variantes de taille quelconque. Prenons l'exemple suivant : *Levitae autem in tribu* ***familiarum suarum*** *non sunt numerati cum eis*. Le vecteur de contextes de ***familiarum suarum*** comprendra tous les n-grammes qui l'entourent :

— Unigrammes : *Levitae, autem, in, tribu, non, sunt, numerati, cum, eis*
— Bigrammes : *Levitae autem, autem in, in tribu, non sunt, sunt numerati, numerati cum, cum eis*
— Trigrammes : *Levitae autem in, autem in tribu, non sunt numerati, sunt numerati cum, numerati cum eis*
— Quadrigrammes : *Levitae autem in tribu, non sunt numerati cum, sunt numerati cum eis*
— Pentagrammes : *non sunt numerati cum eis*

Une fois les vecteurs de contextes construits, nous effectuons le calcul de mesures d'association, afin de mesurer le degré de la relation contextuelle entre la tête du vecteur (*familiarum suarum* dans l'exemple) et chacun de ses contextes. Trois mesures d'association sont utilisées : l'information mutuelle (IM) (Fano, 1961), le rapport des cotes actualisé (RCA) (Evert, 2005) et le rapport de vraisemblance (RV) (Dunning, 1993). Enfin, pour extraire les candidats, nous mesurons à travers le Cosinus (Salton & Lesk, 1968) la similarité entre le n-gramme source et tous les n-grammes candidats du corpus. Notre adaptation de l'approche par sac de mots prend aussi en compte les n-grammes creux (broken n-grams). Ainsi, en plus des n-grammes déjà cités précédemment, et partant du pentagramme *non sunt numerati cum eis*, nous rajoutons les bigrammes suivants : *non numerati, non cum, non eis, sunt cum, sunt eis, numerati eis*. Ceci en supposant que les unigrammes *sunt, numerati*, et *cum*, aient été absents ou omis du corpus à un moment donné. Cette procédure est répétée pour chaque taille de n-gramme.

## 4.4 Approche par plongements de mots

L'approche par plongements de mots (word embeddings) consiste à représenter une variante par un vecteur de plongements. Ce vecteur est calculé à partir d'une combinaison linéaire des vecteurs de plongements des mots qui composent la variante (Arora *et al.*, 2017). Si nous reprenons l'exemple *familiarum suarum*, son vecteur de plongements sera l'addition des vecteurs de plongements des mots *familiarum* et *suarum*. Après le calcul des vecteurs de plongements de tous les n-grammes du corpus, nous classons les candidats à l'aide de la mesure du cosinus. La formule ci-dessous présente le calcul par plongements de mots :

$$Embedding(A) = \sum_{j=1}^{n} Embedding(w_j) \quad (3)$$

avec $A$ qui représente un n-gramme de mots et $n$ le nombre de mots le constituant. $Embedding(w_j)$ correspond au modèle de plongements de mots utilisé. Le résultat de la formule correspond à un modèle de plongements de mots représentant le n-grammes A par : $Embedding(A)$. Une variante de ce modèle serait d'utiliser une somme pondérée pour chaque mot du n-gramme (Wieting *et al.*, 2016). Dans nos expériences nous utilisons deux modèles pré-entraînés pour le latin qui sont Word2Vec [3] et FastText [4].

# 5 Expériences et résultats

Nous avons utilisé la base de données *Beyond Use* [5] qui permet d'étudier les livres d'heures à partir de leurs textes. Cette base fournit une annotation manuelle de variantes textuelles de *l'Obsecro Te* présentes dans 772 manuscrits. Une prière *Obsecro Te* comporte 49 lignes arbitraires définies dans (Plummer & Clark, 2015). Chaque ligne a été comparée et annotée manuellement à la même ligne de la même prière dans les 771 autres copies. À chaque fois qu'une variante est rencontrée, elle est enregistrée comme nouvelle variante dans la base. De ce processus a résulté un corpus d'environ 21 329 entrées d'apparat et 3 298 entrées distinctes. Partant du principe que les informations concernant le type de prière et la segmentation en lignes ne sont pas connues a priori [6], nous n'utilisons pas cette information d'alignement pour extraire les variantes. L'évaluation est faite sur un jeu de tests que l'on divise en 4 listes distinctes. Chaque liste correspond à une taille de n-grammes. Ainsi, nous obtenons une première liste d'unigrammes qui ont comme variantes uniquement des unigrammes, une liste de bigrammes qui ont comme variantes que des bigrammes et ainsi de suite jusqu'aux quadrigrammes [7]. Pour finir, nous rajoutons une cinquième liste notée $Tout$ et qui englobe les quatre précédentes mais qui contient aussi des couples de variantes de tailles variables (environ 23 %).

| | Taille des n-grammes (Taille de la liste d'évaluation) | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 (208) | | | | 2 (82) | | | | 3 (53) | | | | 4 (28) | | | | Tout (482) | | | |
| Méthodes | P | R | F | MAP | P | R | F | MAP | P | R | F | MAP | P | R | F | MAP | P | R | F | MAP |
| DistEdit | **14.0** | **59.1** | **22.6** | 48.3 | 1.82 | 10.4 | 3.11 | 4.65 | 2.83 | 8.49 | 4.24 | 6.04 | 2.85 | 8.06 | 4.21 | 5.43 | 7.01 | 28.1 | 11.2 | 23.1 |
| Jaccard | 11.4 | 50.8 | 18.7 | 37.9 | 7.80 | **66.0** | 13.9 | 48.7 | **11.3** | **66.0** | **19.3** | 38.2 | 7.85 | 43.0 | 13.2 | 22.8 | 7.12 | 35.7 | 11.8 | 25.3 |
| BoW (IM) | 10.2 | 46.2 | 16.8 | 17.3 | 5.24 | 45.3 | 9.40 | 12.5 | 9.24 | 51.9 | 15.6 | 14.8 | 3.21 | 15.6 | 5.33 | 10.5 | 2.54 | 10.8 | 4.11 | 8.36 |
| BoW (RCA) | 10.1 | 46.2 | 16.7 | 17.1 | 4.87 | 41.6 | 8.73 | 12.3 | 9.05 | 50.1 | 15.3 | 14.5 | 3.21 | 15.6 | 5.33 | 10.5 | 2.54 | 10.9 | 4.12 | 8.39 |
| BoW (RV) | 12.6 | 52.6 | 20.3 | **48.5** | **8.04** | 60.9 | **14.2** | 28.6 | 10.7 | 60.0 | 18.2 | 25.7 | 2.85 | 17.7 | 4.78 | 12.1 | **9.70** | **41.7** | **15.7** | **31.9** |
| Word2Vec | 7.74 | 33.7 | 12.5 | 23.3 | 6.95 | 63.3 | 12.4 | **62.3** | 9.43 | 65.0 | 16.4 | **49.1** | **12.5** | **64.0** | **20.9** | **40.9** | 3.89 | 21.6 | 6.60 | 17.2 |
| FastText | 6.39 | 30.2 | 10.5 | 28.7 | 6.95 | 60.9 | 12.4 | 59.7 | 9.43 | 63.9 | 16.4 | 41.1 | 12.1 | 57.3 | 20.0 | 29.0 | 3.25 | 19.5 | 5.57 | 11.6 |

TABLE 1 – Évaluation des approches état-de-l'art et notre adaptation de l'approche distributionnelle (BoW). Les résultats sont représentés en termes de précision (P), rappel (R) et F-mesure (F) au top 10 ainsi que la précision moyenne (MAP). Nous affichons entre parenthèses pour chaque longueur de n-grammes, la taille de la liste d'évaluation. Par exemple : 1(208) correspond à 208 n-grammes pour lesquels nous cherchons à obtenir des variantes unigrammes.

Le tableau 1 illustre les résultats des différentes approches implémentées. L'approche par distance d'édition présente les meilleurs résultats lorsque les variantes sont des unigrammes. En revanche, ses performances chutent de manière prononcée dès lors qu'il s'agit de n-grammes supérieurs à

3. http://www.cs.cmu.edu/~dbamman/latin.html
4. https://github.com/facebookresearch/fastText/blob/master/docs/crawl-vectors.md
5. http://www6.sewanee.edu/BeyondUse/
6. Dans un cadre applicatif réel nous aurons à disposition une transcription via OCR de livres liturgiques.
7. Nous n'allons pas plus loin car il y a très peu de variantes de taille supérieure à 4.

1. De nombreuses variantes sont des permutations, une opération non appréhendée par la distance d'édition qui est sensible à l'ordre dans lequel apparaissent les éléments d'une variante. L'indice de Jaccard, et même s'il est légèrement moins bon que la distance d'édition, obtient des résultats nettement supérieurs dès lors que l'on passe aux n-grammes supérieurs à 1. Ceci est sans doute dû au fait qu'il ne soit pas sensible au phénomène de permutation. Notre adaptation de l'approche distributionnelle (BoW (RV)) utilisant le rapport de vraisemblance comme mesure d'association montre les meilleurs résultats sur la liste globale ($Tout$). Cette approche est celle qui gère le mieux les couples de variantes de taille variable. Les moins bons résultats de BoW (IM) et de BoW (RCA) montrent que l'information mutuelle (IM) et le rapport des cotes actualisé (RCA) sont moins aptes à capturer l'association des n-grammes dans les vecteurs de contextes. Le manque de données est aussi un facteur qui peut expliquer ces résultats. L'approche par plongements de mots (Word2Vec) montre les meilleurs scores en terme de Map pour les n-grammes supérieurs à 1, ce qui suggère que ce modèle est le plus adapté pour cette configuration (n-grammes > 1). Les moins bons résultats concernant les unigrammes peuvent cependant s'expliquer par le fait que Word2Vec et FastText sont des modèles pré-entraînés et des unigrammes de *l'Obsecro Te* n'y figurent pas. Une combinaison d'approches a été menée mais n'a pas montré d'amélioration significative. Si certains phénomènes peuvent être détectés comme, par exemple, les variantes orthographiques, flexionnelles ou dérivationnelles en utilisant la distance d'édition, les permutations en utilisant l'indice de Jaccard ou encore les substitutions lexicales synonymiques grâce aux approches distributionnelles ou par plongements de mots, d'autres variantes sont plus difficiles à détecter comme les variantes d'expansion ou de réduction, et bien entendu les variantes combinant plusieurs opérations linguistiques. Par ailleurs, nous rencontrons des substitutions lexicales problématiques où certains mots ou séquences de mots sont remplacés par des connecteurs (*et, a, que, de, in...*) comme : *sanctam / et, de filio tuo / a, in omnibus / et in*. L'apparition très fréquente des connecteurs les rend difficiles à modéliser pour un type particulier de variantes.

# 6 Conclusion

Cet article a présenté une première étude de l'extraction de variantes textuelles latines provenant de livres liturgiques datant de la fin du Moyen Âge. Si des résultats intéressants ont été observés, les méthodes mises en œuvre ne permettent pas de distinguer les variantes orthographiques des variantes flexionnelles ou dérivationnelles. De plus, même les méthodes adaptées à la détection de certaines variantes échouent sur des cas problématiques : la substitution synonymique est peu performante pour détecter les substitutions lexicales où les éléments substitués ont des distributions très différentes comme celles mettant en jeu des connecteurs. Aucune méthode ne s'est révélée efficace pour la détection des expansions ou des réductions. Néanmoins, ce travail constitue une première amorce qui appelle à continuer dans cette voie et qui peut aussi servir à d'autres tâches comme la segmentation des livres d'heures et la découverte de nouvelles connaissances issues de ce type de ressources.

# Remerciements

Ce travails'inscrit dans le cadre du projet HORAE (Hours - Recognition, Analysis, Editions) et a bénéficié d'une aide de l'Agence nationale de la recherche portant la référence ANR-17-CE38-0008. Nous tenons tout particulièrement à remercier le professeur Gregory Clark pour avoir mis à disposition les données de l'*Obsecro Te* et l'annotation manuelle des variantes textuelles.

# Références

ARORA S., YINGYU L. & TENGYU M. (2017). A simple but tough to beat baseline for sentence embeddings. In *Proceedings of the 17th International Conference on Learning Representations (ICLR'17)*, p. 1–11.

BELYY A., DUBOVA M. & NEKRASOV D. (2018). Improved evaluation framework for complex plagiarism detection. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 157–162 : Association for Computational Linguistics.

BRLEK A., FRANJIC P. & UZELAC N. (2016). Plagiarism detection using word2vec model. In *Text analysis and retrieval 2016 course project*, p. 4–7.

DAILLE B. (2017). *Term Variation in Specialised Corpora : Characterisation, automatic discovery and applications*, volume 19 of *Terminology and Lexicography Research and Practice*. John Benjamins.

DONDI C. (2016). *Printed Books of Hours from Fifteenth-Century Italy : The Texts, the Books, and the Survival of a Long-Lasting Genre.* Firenze : Leo S. Olschki.

DUNNING T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**(1), 61–74.

EVERT S. (2005). *The statistics of word cooccurrences : word pairs and collocations*. PhD thesis, University of Stuttgart.

FANO R. M. (1961). *Transmission of Information : A Statistical Theory of Communications*. Cambridge, MA, USA : MIT Press.

FIRTH J. R. (1957). A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis (special volume of the Philological Society)*, p. 1–32. Oxford : Blackwell.

HARRIS Z. S. (1971). *Structures mathématiques du langage*. Dunod. Traduit de l'Américain par C. Fuchs.

HE H., GIMPEL K. & LIN J. (2015). Multi-perspective sentence similarity modeling with convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 1576–1586 : Association for Computational Linguistics.

JACCARD P. (1901). Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, **37**, 547–579.

KANHIRANGAT V. & GUPTA D. (2016). Detection of idea plagiarism using syntax - semantic concept extractions with genetic algorithm. *Expert Systems with Applications*, **73**.

KOCIC A. (2008). The problem of synforms. *Facta Universitatis*, **6**(1), 51–59.

LAUFER B. (1988). The concept of 'synforms' (similar lexical forms) in vocabulary acquisition. *Language and Education*, **2**(2), 113–132.

LE H., N. PHAM L., D. NGUYEN D., V. NGUYEN S. & N. NGUYEN A. (2016). Semantic text alignment based on topic modeling. p. 67–72.

LEROQUAIS V. (1927). *Les livres d'heures manuscrits de la Bibliothèque nationale*. Paris.

LEVENSHTEIN V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, **10(8)**, 707–710.

LÉONELLI M.-C. (1985). La dévotion aux saints d'après les livres d'heures confectionnés à avignon. In *Mémoires de l'académie de Vaucluse*, volume 6, p. 327–335.

PLUMMER J. & CLARK G. T. (2015). Obsecro te. *Beyond Use : A Digital Database of Variant Readings In Late Medieval Books of Hours*. http://www6.sewanee.edu/BeyondUse/texts_list.php\?texts=ObsecroTe.

SALTON G. & LESK M. E. (1968). Computer evaluation of indexing and text processing. *Journal of the Association for Computational Machinery*, **15**(1), 8–36.

SANCHEZ-PEREZ M. A., GELBUKH A. F., SIDOROV G. & GÓMEZ-ADORNO H. (2018). Plagiarism detection with genetic-based parameter tuning. *IJPRAI*, **32**(1), 1–23.

STUTZMANN D. (2015). Les écritures des livres d'heures dans l'espace français (1290-1550). In *Proceedings of the 19th Colloquium of the Comité international de paléographie latine*.

WIETING J., BANSAL M., GIMPEL K. & LIVESCU K. (2016). Towards universal paraphrastic sentence embeddings. *International Conference on Learning Representations, CoRR*, **abs/1511.08198**.

# Transformation d'annotations en parties du discours et lemmes vers le format Universal Dependencies : étude de cas pour l'alsacien et l'occitan

Aleksandra Miletić[1] Delphine Bernhard[2] Myriam Bras[1]
Anne-Laure Ligozat[3] Marianne Vergez-Couret[4]
(1) CLLE-ERSS, CNRS, Université Toulouse-Jean Jaurès
(2) LiLPa - EA 1339, Université de Strasbourg
(3) LIMSI, CNRS, ENSIIE, Université Paris-Saclay, F-91405 Orsay
(4) FoReLLIS - EA 3816, Université de Poitiers
aleksandra.miletic@univ-tlse2.fr, dbernhard@unistra.fr,
myriam.bras@univ-tlse2.fr, annlor@limsi.fr,
marianne.vergez.couret@univ-poitiers.fr

RÉSUMÉ

Cet article présente un retour d'expérience sur la transformation de corpus annotés pour l'alsacien et l'occitan vers le format CONLL-U défini dans le projet *Universal Dependencies*. Il met en particulier l'accent sur divers points de vigilance à prendre en compte, concernant la tokénisation et la définition des catégories pour l'annotation.

ABSTRACT

**Converting POS-tag and Lemma Annotations into the Universal Dependencies Format : A Case Study on Alsatian and Occitan**

This article presents a retrospective report on the transformation of annotated corpora for Alsatian and Occitan into the CONLL-U format defined in the Universal Dependencies project. In particular, it emphasizes various issues to be taken into account, concerning the tokenization and the definition of the categories.

MOTS-CLÉS : annotation, alsacien, occitan, Universal Dependencies.

KEYWORDS: annotation, Alsatian, Occitan, Universal Dependencies.

## 1 Introduction

Les langues régionales de France sont à l'heure actuelle encore largement sous-dotées en ressources linguistiques, qu'il s'agisse de corpus, bruts ou annotés, ou de ressources lexicales comme des lexiques flexionnels ou des dictionnaires bilingues. Nous nous intéressons dans cet article à l'alsacien et à l'occitan qui, bien que se trouvant dans des situations différentes (famille linguistique, vitalité, ressources existantes), font face à des défis communs. En effet, le développement de ressources et d'outils pour les langues peu dotées nécessite une approche pragmatique qui prend en compte les faibles moyens financiers et humains généralement disponibles. Ainsi, Soria *et al.* (2013) ont énoncé un ensemble de principes mettant l'accent sur la coopération, l'utilisation de standards internationaux, la réutilisation d'approches et de ressources existantes, le partage des ressources et outils produits dans des formats ouverts.

Dans cet article nous montrons comment nous avons repris ces divers principes à notre compte pour le

développement de corpus annotés en parties du discours (POS) et lemmes dans le format CONLL-U [1] défini dans le projet *Universal Dependencies* (UD) (Nivre *et al.*, 2016), pour l'alsacien et l'occitan. Le choix du format CONLL-U répond à au moins deux des principes énoncés par Soria *et al.* (2013) : utilisation de standards et réutilisation d'approches existantes. Par ailleurs, les dialectes alsaciens et l'occitan ne sont pour l'heure pas représentés dans les corpus *Universal Dependencies*. Nous souhaitons pouvoir combler ce manque en partageant les ressources annotées produites avec une licence libre. Cet objectif respecte deux autres principes de Soria *et al.* (2013), à savoir le partage des ressources sur une plateforme pérenne et avec une licence libre, et permet en outre de donner une meilleure visibilité à ces langues. Par ailleurs, le format CONLL-U peut être utilisé directement pour entraîner divers outils (par exemple, spaCy [2] ou UDPipe (Straka & Straková, 2017)), ce qui constitue également un atout pour les langues disposant de peu de moyens pour développer de nouveaux outils.

Nous mettons aussi l'accent sur divers points de vigilance qui sont à prendre en compte lors du passage de corpus annotés vers le format CONNL-U et le jeu d'étiquettes morphosyntaxiques définies dans le projet *Universal Dependencies*. Nous souhaitons ainsi que ce retour d'expérience puisse servir pour d'autres langues, et notamment des langues peu dotées. Même si ce travail a déjà été réalisé pour d'autres langues mieux dotées, cela ne le rend pas trivial pour autant, et ce d'autant plus que les comptes rendus de ce type d'expériences ont peu été publiés.

# 2 Annotation de corpus au format UD

Le projet *Universal Dependencies* vise à proposer des corpus annotés de manière cohérente pour différentes langues, grâce à des principes d'annotation communs et des catégories unifiées pour les parties du discours, les propriétés lexicales et grammaticales des mots, et les relations de dépendance syntaxique (Nivre *et al.*, 2016). Depuis les origines du projet, le nombre de corpus arborés et de langues répertoriées ne cesse de croître. On trouve notamment des langues considérées comme peu dotées, représentées par des corpus de petite taille (quelques milliers de tokens) comme le breton, le féroïen ou encore le komi-zyriène.

Certains corpus sont directement annotés en suivant les catégories pré-définies dans *Universal Dependencies*, comme par exemple le corpus komi-zyriène produit par le Lattice pour des besoins d'évaluation (Lim *et al.*, 2018). Cependant, d'autres standards existent, comme par exemple le standard GRACE (Rajman *et al.*, 1997), lui-même dérivé des jeux d'étiquettes MULTEXT (Ide & Véronis, 1994) et EAGLES (von Rekowski, 1996). Ils ont été très largement employés pour l'annotation de multiples corpus dans plusieurs langues, dont l'occitan, l'une des langues étudiées dans cet article (cf. section 3).

De nombreux corpus UD ont ainsi été produits par transformation semi-automatique à partir de corpus étiquetés et arborés existants, à l'aide notamment de tables de conversion entre jeux d'étiquettes [3]. Cela étant, ce processus de conversion peut être plus complexe et nécessiter des opérations supplémentaires. Nous nous intéressons ici plus particulièrement aux étapes de découpage en tokens et à l'annotation en lemmes et parties du discours qui concernent directement les travaux présentés dans cet article. Concernant ces aspects, Chun *et al.* (2018) décrivent le processus pour trois corpus arborés du coréen. La transformation vers le format UD version 2.0 a nécessité un redécoupage en tokens de l'un des corpus, pour les tokens incluant des signes de ponctuation et des symboles. Pour les deux autres corpus, des tables de correspondance ont été établies entre les étiquettes morphosyntaxiques

1. `http://universaldependencies.org/docs/format.html`
2. `https://spacy.io/`
3. Voir `https://universaldependencies.org/tagset-conversion/index.html`

existantes et celles du projet UD. Sanguinetti *et al.* (2018) décrivent le processus de conversion vers UD d'un corpus de tweets en italien. Certains tokens ont dû être redécoupés, à savoir les contractions préposition-article et verbe-clitique et un certain nombre de lemmes ont été ajoutés manuellement, car ne pouvant faire l'objet d'une lemmatisation automatique. Les formes non standards (mots étrangers ou dialectaux, formes tronquées ou amalgamées) ont été étiquetées avec la catégorie `X`.

Les objectifs d'annotation unifiée à travers les langues du projet UD posent bien évidemment divers problèmes qui ont été soulignés à plusieurs reprises. L'annotation en relations de dépendance syntaxique impose un découpage particulier en tokens (voir (Gerdes & Kahane, 2016) pour une discussion de ces choix). Certains mots *orthographiques*, tels que produits généralement par les outils de tokénisation, doivent être découpés en plusieurs unités. C'est notamment le cas pour les formes contractées ou des formes non standard, que l'on peut trouver dans les contenus produits par les utilisateurs sur le web (Alonso *et al.*, 2016). Ce découpage particulier en tokens a d'ailleurs été un des problèmes auxquels nous avons été confrontées (voir section 4). Par contre, l'annotation en lemmes et parties du discours est moins problématique en raison de la flexibilité du format. Nous avons tout de même choisi de ne pas respecter systématiquement les définitions proposées pour chaque partie du discours en raison des spécificités des langues traitées (voir section 5.3).

# 3 Jeux d'étiquettes originaux pour l'alsacien et l'occitan

Les jeux d'étiquettes originaux relèvent de deux situations différentes : celui utilisé pour les dialectes alsaciens était déjà très proche des *Universal POS tags* car s'en inspirant fortement, tandis que celui de l'occitan s'inspirait des catégories plus nombreuses de GRACE (Rajman *et al.*, 1997). En plus des parties du discours, d'autres informations ont également été ajoutées lors de l'annotation : le lemme et sa traduction en français, ainsi que les noms de lieux. Ces annotations supplémentaires permettent de constituer directement des lexiques bilingues à partir des corpus, et d'évaluer l'annotation en entités de type lieu. Les corpus annotés sont de taille sensiblement similaire pour les deux langues : environ 12 600 tokens pour l'alsacien et 11 900 pour l'occitan (Bernhard *et al.*, 2018c).

Les catégories initiales pour l'alsacien sont très proches des catégories proposées dans UD, respectant ainsi le principe de réutilisation soutenu par Soria *et al.* (2013). Il existe peu de descriptions linguistiques détaillées de la morphosyntaxe de l'alsacien et il nous semblait donc plus réaliste de partir d'un nombre limité de catégories bien décrites, compte tenu également du temps limité et des moyens à disposition. Nous avons toutefois ajouté les catégories suivantes aux 17 catégories UD (Bernhard *et al.*, 2018b) : `EPE` pour les épenthèses (insertions de sons pour faciliter l'articulation), `APPRART` pour les contractions préposition + article, `MOD` pour les modaux et `FM` pour les mots d'une autre langue (souvent le français). Le cas des épenthèses est discuté dans la section 5.2 et celui des contractions le sera dans la section suivante. Les catégories `MOD` et `FM` ont été ajoutées car relativement claires et simples à annoter à ce stade.

Pour l'occitan, deux ressources ont été produites : Loflòc, lexique de formes fléchies (Vergez-Couret, 2016; Bras *et al.*, 2017) puis un corpus annoté en catégories morphosyntaxiques, avec un même jeu d'étiquettes, adapté du standard GRACE (Bras *et al.*, 2018). Les étiquettes GRACE peuvent être interprétées comme des étiquettes à 3 niveaux. Le premier niveau d'étiquette permet d'indiquer la catégorie grammaticale des formes fléchies, de classer les signes de ponctuation (`F`) et les formes attestées dont la classification n'a pas encore été réalisée (`X`). Le deuxième niveau propose une classification sémantique ou fonctionnelle spécifique à chaque catégorie de niveau 1. Le troisième niveau concerne les informations morphosyntaxiques flexionnelles, telles le genre, le nombre, la personne, le temps verbal, etc. Des modifications ont été apportées par rapport au jeu d'étiquettes

GRACE. Par exemple, l'ajout, pour les verbes, d'un attribut "Form" pouvant prendre les valeurs positif/négatif afin d'annoter l'impératif négatif de l'occitan dont la forme se distingue de celle de l'impératif positif. Une description détaillée du jeu d'étiquettes est disponible dans le guide d'annotation (Bras, 2018).

| **tokens** | Cossí | aquò | pòt | èsser | ? |
|---|---|---|---|---|---|
| **VO** | Rx | Pd | Vm | Vm | F |
| **UD** | ADV | PRON | VERB | VERB | PUNCT |
| **lemme** | cossí | aquò | poder | èsser | ? |
| **glose** | comment | ça | pouvoir | être | ? |

FIGURE 1 – Phrase annotée en occitan. VO = étiquettes originales, UD = étiquettes après conversion.

Dans la version actuelle du corpus, nous retenons les deux premiers niveaux d'étiquettes (le POS et la sous-catégorie sémantique ou fonctionnelle, cf. Figure 1), ce qui donne un jeu de 40 étiquettes au total ; ces annotations ont été corrigées manuellement après une première annotation automatique. Le corpus a également été étiqueté en informations flexionnelles, mais cette couche d'information, produite de manière automatique, n'a pas encore été validée par des annotateurs humains. Ceci fait partie des pistes pour la suite du travail. Le choix de ce jeu d'étiquettes est principalement motivé par le fait d'exploiter des ressources existantes pour l'occitan (le lexique mentionné ci-dessus), mais aussi pour d'autres langues proches, notamment pour le français (cf. FTB (Abeillé *et al.*, 2003), construit en utilisant les mêmes standards) et pour le catalan (cf. AnCora-CA (Taulé *et al.*, 2008), fondé sur des principes similaires).

Pour rendre nos corpus compatibles avec les exigences du projet UD, il a été nécessaire d'intervenir à deux niveaux : il a fallu adapter le découpage en tokens (cf. section 4) et ensuite convertir l'annotation morphosyntaxique vers le jeu d'étiquettes UD (cf. section 5). Le travail s'est ici partagé entre les deux équipes spécialistes de l'alsacien et de l'occitan, pour les questions plus linguistiques, et la réalisation pratique a été supervisée par le LIMSI concernant la mise en oeuvre des scripts de conversion et de vérification. Ce partage du travail a permis une mise en commun des solutions aux problèmes rencontrés, garantissant ainsi une meilleure qualité aux ressources produites.

# 4 Transformation vers UD : segmentation en tokens

Comme il a été mentionné dans la section 2, le guide de tokénisation UD pose des exigences spécifiques [4]. Tout d'abord, il considère comme unité textuelle de base un token *syntaxique* et non pas *orthographique*. Cela implique, par exemple, la séparation des formes contractées *préposition + article*, présentes aussi bien en alsacien qu'en occitan. En même temps, le projet proscrit les mots « multi-tokens » (incluant des espaces) et les expressions polylexicales (*multiword expressions*) sont systématiquement traitées par l'annotation syntaxique plutôt que par la tokénisation, sauf dans quelques cas exceptionnels.

## 4.1 Formes contractées

Le découpage d'un texte en tokens syntaxiques n'est pas une question anodine. Gerdes & Kahane (2016) notent que cette approche favorise le principe d'adéquation de l'annotation linguistique, mais qu'elle est contraire au principe de simplicité, vu qu'elle peut entraîner le besoin d'une validation manuelle de la tokénisation. Et dans le cadre d'une conversion d'un corpus existant, il faut non seulement opérer les découpages, mais aussi fournir une annotation pour les formes obtenues.

4. Voir `https://universaldependencies.org/u/overview/tokenization.html`

Même dans le cas d'une langue avec un nombre limité de ces formes, ce passage peut se montrer problématique. En décrivant la conversion du corpus catalan, Alonso & Zeman (2016) indiquent que le catalan dispose de 6 formes contractées (*al*, *als*, *del*, *dels*, *pel*, *pels*) et décrivent leur découpage et leur ré-annotation. Or, dans le corpus distribué (v2.3), on retrouve un nombre élevé de ces formes non traitées (au-delà de 14 000 dans l'ensemble du corpus). En revanche, la séparation des clitiques semble bien effectuée (cf. *ofrir-los* traité comme deux formes séparées, *ofrir* et *los*). L'occitan languedocien dispose des mêmes formes contractées que le catalan, plus les formes *sul*, *suls*, *jol*, *jols*, *vèl* et *vèls*. Pour les autres dialectes, les formes contractées diffèrent légèrement mais il est également possible de les lister. Elles portent des étiquettes spéciales concaténées (`SpDa`), ce qui facilite leur repérage et l'identification des étiquettes à attribuer aux formes découpées (*al* `SpDa` → *a* `Sp` + *lo* `Da`).

Pour l'alsacien en revanche, il existe une grande variabilité dans les graphies de ces formes : on trouve par exemple les variantes *zuem*, *züem*, et *zum* pour *zu* + *dem*. Il est donc difficile d'en établir une liste *a priori* et, par conséquent, de les identifier lors de la tokénisation d'un texte brut. Comme pour l'occitan, nous disposons d'une étiquette POS dédiée, ce qui facilitait le repérage des formes contractées. En effet, le jeu d'étiquettes initial intègre la catégorie `APPRART` pour les contractions préposition + article. Cette catégorie est directement tirée du jeu d'étiquettes STTS pour l'allemand (Schiller *et al.*, 1999). Le guide d'annotation STTS nous a été très utile pour constituer notre propre guide et résoudre des difficultés rencontrées lors de l'annotation manuelle. Nous avons procédé à un découpage semi-automatique des tokens annotés `APPRART` dans le corpus alsacien, afin de les segmenter en deux tokens, l'un étiqueté `ADP`, l'autre `DET`. En raison de la variation graphique, 40 formes de ce type ont été repérées dans le corpus. La Figure 2 donne l'exemple d'une phrase annotée en alsacien. Dans cet exemple, *Mìtem* a été segmenté en *Mìt* + *dem*.

| **tokens** | Mìtem | | Sabayon | ìwwerzìehje | ùn | mìt | de | g'hobelte | Màndle | bstraie | . |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **VO** | APPRART | | NOUN | VERB | CONJ | ADP | DET | ADJ | NOUN | VERB | PUNCT |
| **UD** | ADP | DET | NOUN | VERB | CCONJ | ADP | DET | ADJ | NOUN | VERB | PUNCT |
| **lemme** | mìt | de | Sabayon | ìwwerzìehje | ùn | mìt | de | g'hobelt | Màndel | bstraie | . |
| **glose** | avec | le | sabayon | napper | et | avec | les | effilé | amande | saupoudrer | . |

FIGURE 2 – Phrase annotée en alsacien. VO - étiquettes originales, UD - étiquettes après conversion.

## 4.2 Traitement des mots « multi-tokens »

L'exigence d'UD de ne pas utiliser de mots « multi-tokens » ne concerne que le corpus occitan. À la différence du corpus alsacien, découpé en tokens orthographiques, le corpus occitan contient un certain nombre d'unités polylexicales qui ont été soudées en un seul token lors de l'annotation manuelle. Il s'agit notamment de séquences figées qui n'ont pas de lecture libre, le plus souvent des locutions adverbiales ou conjonctives qui peuvent s'écrire aussi en un seul mot graphique (*ça_que_la* 'pourtant', *si_que_non* 'sinon'). La liste complète de ces formes est disponible dans le guide d'annotation (Bras, 2018). Le repérage et le découpage automatique de ces formes n'est pas problématique, vu l'utilisation systématique du caractère "_". En revanche, l'attribution des étiquettes aux formes découpées peut l'être : certains de ces figements ne sont pas transparents au niveau de leur structure syntaxique et l'identification des parties du discours de leurs éléments n'est pas intuitive, même pour un annotateur humain (cf. la forme *ça_que_la*, annotée PRON SCONJ ADV, où la reconnaissance de *la* comme adverbe de lieu passe par le repérage, pour l'annotateur, de la forme plus fréquente *lai*). Dans la version actuelle du corpus, ces unités se présentent encore comme tokens multi-mots. Elles seront traitées dans l'étape suivante du travail, avant d'aborder l'annotation syntaxique.

# 5 Transformation vers UD : étiquettes de parties du discours

Les corpus n'ayant pas été annotés dans le format UD dès le départ, il a été nécessaire de convertir les étiquettes initiales. Cette tâche était moins exigeante pour l'alsacien, dont le jeu d'étiquettes original est majoritairement fondé sur celui de UD (cf. section 3). La transformation du jeu d'étiquettes occitan a posé plus de défis, notamment dus à des différences de granularité et de découpage des catégories.

## 5.1 Définition d'une table de correspondances

Pour la conversion, nous avons utilisé des tables de correspondance dans des scripts de conversion automatique. La difficulté principale a été de définir ces correspondances en prenant en compte les caractéristiques des langues traitées. Le passage vers le jeu d'étiquettes UD nous a permis de vérifier la cohérence des annotations des corpus initiaux (par exemple, cohérence des lemmes pour les nombres en occitan). La détection de ces incohérences a été effectuée de manière automatique et la correction a ensuite été réalisée manuellement.

Pour l'alsacien, nous avons conservé les étiquettes initiales, pour celles se trouvant dans le jeu UD. La catégorie `MOD` devient `AUX`, `FM` devient `X`. Les cas de `APPRART` et `EPE` sont discutés respectivement dans les sections 4.1 et 5.2.

Pour l'occitan, certaines étiquettes avaient des correspondants du même niveau de granularité dans le jeu UD (nom commun : `Nc` → `NOUN`; nom propre : `Np` → `PROPN`; verbe principal : `Vm` → `VERB`; verbe auxiliaire : `Va` → `AUX`). Mais ce passage d'un jeu de 40 étiquettes vers un jeu de 17 étiquettes a également entraîné une perte d'information dans l'étiquette UD (mais pas dans le corpus, qui conserve également l'étiquette d'origine). Différents types d'adverbes (`Rg` - adv. général, `Rx` - adv. interrogatif/exclamatif, `Rq` - adv. d'intensité/de quantité) sont tous exprimés par une seule étiquette - `ADV`. Sur les 8 étiquettes pronominales de départ, 7 sont traduites par `PRON`. La conversion a également soulevé des questions de découpage des catégories : dans le jeu original, les cardinaux sont annotés en fonction de leur comportement syntaxique comme noms, adjectifs, pronoms ou déterminants cardinaux. Or, dans le cadre de UD, les numéraux sont traités comme une seule catégorie – `NUM`.

Néanmoins, les pertes d'information décrites ci-dessus peuvent être neutralisées si l'on adopte également l'annotation en traits lexicaux et flexionnels. Ceci permettrait de rétablir les distinctions entre les étiquettes pronominales : les pronoms personnels, démonstratifs, indéfinis, relatifs et interrogatifs porteraient différentes valeurs du trait `PronType` (`PronType=Prs|Dem|Ind|Rel|Int`), alors que les pronoms possessif et réflexif seraient marqués comme `PronType=Prs`, mais ils porteraient également des traits supplémentaires, respectivement `Poss=Yes` et `Reflex=Yes`.

## 5.2 Cas de l'épenthèse

Pour le corpus alsacien, nous avons choisi d'annoter les épenthèses avec une étiquette EPE : *fànga␣**-n-/EPE**␣à drucka*. Ce phénomène a été annoté de diverses manières dans d'autres corpus existants : dans la version UD du corpus français Sequoia (Candito & Seddah, 2012), le "-t-" de liaison (dans "semble-t-il" par exemple) est annoté comme PART ; le guide d'annotation du corpus TCOF-POS (Benzitoun *et al.*, 2012) prévoit une étiquette EPE mais elle n'est utilisée qu'une fois dans le corpus ("que **l'/EPE** on ne voit"). Dans notre cas particulier, EPE a été transformé en PART.

## 5.3 Différences dans les définitions

D'une manière générale, l'existence de catégories qui semblent identiques dans la liste des catégories initiales et la liste UD ne signifie pas nécessairement que ces catégories ont la même définition. Par

exemple, en alsacien, nous avons annoté les particules verbales séparées PART, ce qui ne correspond pas aux principes UD qui recommandent d'étiqueter les particules séparables ADP ou ADV en fonction de leur type d'origine [5]. Nous avons dans ce cas choisi de nous référer aux recommandation STTS pour la catégorie PTKVZ (*abgetrennter Verbzusatz*, classé dans la catégorie *Partikel* - particule). Dans ce cas bien précis, la transformation vers UD voudrait que soit vérifié chaque token étiqueté PART de manière à choisir soit ADP, soit ADV. Ce travail n'a pour l'heure pas été effectué.

Pour l'occitan, l'expression de la possession dans le groupe nominal pouvant être réalisée par le déterminant *mon* (cf. *mon filh*) ou l'adjectif *miu* (cf. *lo mieu filh* ou *lo filh mieu*), l'étiquette As (adj. possessif) a été traduite par ADJ et non par DET, contrairement aux préconisations du guide UD.

### 5.4 Informations additionnelles

Enfin, les informations additionnelles produites lors de l'annotation sont ajoutées dans la dernière colonne du fichier CONLL-U (glose en français et informations sur les entités de lieux). Comme prévu dans le format CONLL-U, nous avons également conservé l'annotation d'origine dans une colonne.

## 6 Conclusion et perspectives

Nous avons présenté un retour d'expérience sur la transformation de deux corpus annotés en alsacien et occitan vers le format UD. Notre travail de réflexion sur l'annotation a démarré au moment de la publication de UD v1 (Nivre *et al.*, 2016). Entre temps, le projet UD a pris de l'ampleur, et une version 2 des recommandations a été publiée. Il nous a donc semblé opportun de transformer nos corpus annotés vers le format préconisé par UD, pour des questions de visibilité et de partage de nos ressources, ainsi que pour aborder l'étape suivante de l'analyse syntaxique en s'appuyant sur les travaux réalisés dans des langues proches dans le cadre de UD. Comme nous l'avons montré dans l'article, une telle transformation peut poser divers problèmes, qui sont plus ou moins simples à résoudre. Il faut tout d'abord souligner le travail de réflexion qui est nécessaire en amont de la transformation. La transformation des corpus d'origine a été réalisée de manière essentiellement automatique et semi-automatique pour partie (nouveau découpage en tokens notamment). Ce travail sur des langues peu dotées n'aurait pu être réalisé sans une réelle coopération entre diverses équipes, dotées de compétences complémentaires, ce qui permet de gagner en efficacité. En effet, le travail en parallèle sur plusieurs langues a permis de profiter des expériences réalisées sur d'autres langues. Les problèmes qui se posent très vite sur une langue permettent une vigilance accrue à ce sujet dans une autre langue. Par ailleurs, les outils développés peuvent être réutilisés (scripts de conversion et de vérification notamment). Une première version de nos corpus au format CONLL-U est disponible (Bras *et al.*, 2018; Bernhard *et al.*, 2018a). Dans l'avenir, les informations manquantes seront également ajoutées : informations sur les propriétés lexicales et grammaticales des mots et relations syntaxiques. Pour l'occitan, l'annotation en dépendances est en cours dans le cadre du projet Linguatec.

## Remerciements

Les travaux décrits dans cet article ont bénéficié du soutien de l'ANR (projet RESTAURE - référence ANR-14-CE24-0003).

5. http://universaldependencies.org/u/pos/PART.html

# Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for French. In *Treebanks*, p. 165–187. Springer.

ALONSO H. M., SEDDAH D. & SAGOT B. (2016). From Noisy Questions to Minecraft Texts : Annotation Challenges in Extreme Syntax Scenario. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, p. 13–23.

ALONSO H. M. & ZEMAN D. (2016). Universal Dependencies for the AnCora treebanks. *Procesamiento del Lenguaje Natural*, (57), 91–98.

BENZITOUN C., FORT K. & SAGOT B. (2012). TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe. In *JEP-TALN 2012 - Journées d'Études sur la Parole et conférence annuelle du Traitement Automatique des Langues Naturelles*, p. 99–112, Grenoble, France.

BERNHARD D., ERHART P., HUCK D. & STEIBLÉ L. (2018a). Annotated Corpus for the Alsatian Dialects. 10.5281/zenodo.2536041.

BERNHARD D., ERHART P., HUCK D. & STEIBLÉ L. (2018b). *Part-of-Speech Annotation Guidelines for the Alsatian Dialects*. 10.5281/zenodo.1171925.

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. (2018c). 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.

BRAS M. (2018). *Part-of-Speech Annotation Guidelines for the Occitan Language*. Rapport interne, UMR 5263 CLLE-ERSS, University of Toulouse. 10.5281/zenodo.1182949.

BRAS M., ESHER L., SIBILLE J. & VERGEZ-COURET M. (2018). *Annotated Corpus for Occitan*. Rapport interne, UMR 5263 CLLE-ERSS, University of Toulouse. 10.5281/zenodo.1182949.

BRAS M., VERGEZ-COURET M., HATHOUT N., SIBILLE J., SÉGUIER A. & DAZÉAS B. (2017). Loflòc : Lexic obert flechit occitan. In *XIIème Congrès de l'Association Internationale d'Etudes Occitanes*, Albi, France : Jean-François Courouau et al.

CANDITO M. & SEDDAH D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Actes de TALN 2012 - 19e conférence sur le Traitement Automatique des Langues Naturelles*, p. 321–334.

CHUN J., HAN N.-R., HWANG J. D. & CHOI J. D. (2018). Building Universal Dependency Treebanks in Korean. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

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

IDE N. & VÉRONIS J. (1994). Multext (Multilingual Tools and Corpora). In *14th Conference on Computational Linguistics (COLING'94), Kyoto, Japan*.

LIM K., PARTANEN N. & POIBEAU T. (2018). Analyse syntaxique de langues faiblement dotées à partir de plongements de mots multilingues. *Traitement Automatique des Langues*, **59**(3).

NIVRE J., DE MARNEFFE M.-C., GINTER F., GOLDBERG Y., HAJIC J., MANNING C. D., MCDONALD R., PETROV S., PYYSALO S., SILVEIRA N. & OTHERS (2016). Universal dependencies

v1 : A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

RAJMAN M., LECOMTE J. & PAROUBEK P. (1997). *Format de description lexicale pour le français. Partie 2 : Description morpho-syntaxique*. Rapport interne, EPFL & INaLF. GRACE GTR-3-2.1.

SANGUINETTI M., BOSCO C., LAVELLI A., MAZZEI A., ANTONELLI O. & TAMBURINI F. (2018). PoSTWITA-UD : an Italian Twitter Treebank in Universal Dependencies. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

SCHILLER A., TEUFEL S., STÖCKERT C. & THIELEN C. (1999). *Guidelines für das Tagging deutscher Textcorpora mit STTS*. Rapport interne, Universität Stuttgart & Universität Tübingen.

SORIA C., MARIANI J. & ZOLI C. (2013). Dwarfs sitting on the giants' shoulders–how LTs for regional and minority languages can benefit from piggybacking major languages. In *Proceedings of XVII FEL Conference*, p. 73–79.

STRAKA M. & STRAKOVÁ J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, p. 88–99, Vancouver, Canada.

TAULÉ M., MARTÍ M. A. & RECASENS M. (2008). Ancora : Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the 6th International Conference on Language Ressources and Evaluation (LREC2008)*, Marrakech, Morocco.

VERGEZ-COURET M. (2016). *Description du lexique Loflòc*. Research report, CLLE-ERSS.

VON REKOWSKI U. (1996). *ELM-FR : A typed French incarnation of the EAGLES-TS – Definition of Lexical Specification and Classification Guidelines*. Rapport interne, GSI-Erli.

# Un corpus libre, évolutif et versionné en entités nommées du français

Yoann Dupont
LIFO, université d'Orléans, 6 rue Léonard de Vinci BP 6759, 45067 Orléans cedex 2
yoann.dupont@univ-orleans.fr

RÉSUMÉ

Les corpus annotés sont des ressources difficiles à créer en raison du grand effort humain qu'elles impliquent. Une fois rendues disponibles, elles sont difficilement modifiables et tendent à ne pas évoluer pas dans le temps. Dans cet article, nous présentons un corpus annoté pour la reconnaissance des entités nommées libre et évolutif en utilisant les textes d'articles Wikinews français de 2016 à 2018, pour un total de 1191 articles annotés. Nous décrivons succinctement le guide d'annotation avant de situer notre corpus par rapport à d'autres corpus déjà existants. Nous donnerons également un accord intra-annotateur afin de donner un indice de stabilité des annotations ainsi que le processus global pour poursuivre les travaux d'enrichissement du corpus.

ABSTRACT

**A free, evolving and versioned french named entity recognition corpus.**

Annotated corpora are very hard resources to make because of the high human cost they imply. Once released, they are hardly modifiable and tend to not evolve through time. In this article we present a free and evolving corpus annotated in named entity recognition based on French Wikinews articles from 2016 to 2018, for a total of 1191 articles. We will briefly describe the annotation guidelines before comparing our corpus to various corpora of comparable nature. We will also give an intra-annotator-agreement to provide an estimation of the stability of the annotation as well as the overall process to develop the corpus.

MOTS-CLÉS : reconnaissance des entités nommées, annotation manuelle, corpus annoté.

KEYWORDS: named entity recognition, manual annotation, annotated corpus.

## 1 Introduction

La reconnaissance des entités nommées est une tâche importante du TAL. Elle permet *« l'accès à l'information »* (Nouvel *et al.*, 2015) pour d'autres tâches, comme par exemple la construction d'une base de connaissances (Surdeanu, 2013) ou les systèmes de questions-réponses (Han *et al.*, 2017). La notion d'entité nommée a évolué avec le temps. Tout d'abord considérées comme *« tous les noms propres et quantités d'intérêt »* dans la campagne MUC-6 (Grishman & Sundheim, 1996), où les entités cibles étaient les personnes, lieux, organisations, temps et pourcentages, leur reconnaissance devait aider au remplissage automatique de formulaire. La campagne ACE (Doddington *et al.*, 2004) a donné comme périmètre aux entités nommées les personnes, les organisations, les lieux, les bâtiments, les armes, les véhicules et les entités géo-politiques, dans le cadre de la détection d'événements. Sekine & Nobata (2004), quant à eux, définissent 150 types d'entités nommées organisés de façon

hiérarchique, afin de couvrir un maximum de cas d'utilisation. Ils préconisaient d'élaguer la hiérarchie afin de correspondre au mieux au cas d'usage particuliers. Grouin *et al.* (2011) proposent également une définition d'entités nommées généraliste et couvrante. Ils définissent, en plus des types d'entités, leurs *composants* ainsi que leur structuration. Les entités nommées ont un caractère référentiel et, comme nous venons de le voir, ont également une visée applicative très concrète et sont fortement liées à leur corpus. Un définition rendant compte de toutes ces caractéristiques est celle d'Ehrmann (2008) : *« étant donnés un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus. »*.

Un inconvénient des corpus actuellement disponibles est leur côté figé dans le temps. En effet, le FTB (Abeillé *et al.*, 2003) annoté en entités nommées (Sagot *et al.*, 2012) contient des phrases extraites d'articles du Monde de 1989 à 1995, la partie anglaise du corpus CoNLL 2003 (Tjong Kim Sang & De Meulder, 2003) est un extrait du corpus Reuters d'Août 1996 à Août 1997, la partie allemande comprend, elle, des textes de l'année 1992. Des corpus sur des données plus récentes existent, comme par exemple le corpus d'oral transcrit ESTER2 (Galliano *et al.*, 2009) qui couvre les années 1999 à 2003, dont les transcriptions ont été utilisées pour le corpus Quaero (Galibert *et al.*, 2010; Rosset *et al.*, 2012). Il nous parait important de diffuser des corpus dont les sources sont aussi récentes que possible, pour plusieurs raisons. La première est d'élargir l'éventail des entités couvertes dans les corpus que nous avons à notre disposition, en effet, de nombreuses nouvelles entités nommées apparaissent régulièrement dans les textes journalistiques et il est important de pouvoir les attester. Un autre intérêt, qui découle du précédent, est que cela permet d'évaluer les systèmes construits pour ces tâches particulières et d'évaluer dans des conditions réalistes leur qualité et leur robustesse.

Au-delà des divergences définitoires, comme le fait d'annoter ou non le titre d'une personne, les corpus annotés ne sont pas exempts d'erreurs. Il existe divers articles les évoquant ou les corrigeant, nous pouvons citer Finkel *et al.* (2004) pour le corpus GENIA (Kim *et al.*, 2003) et Nouvel *et al.* (2010) pour ESTER2. Pourtant, il a été montré que les erreurs humaines étaient une source d'erreur pour les systèmes à base d'apprentissage (Boisen *et al.*, 2000). Pour ces raisons, il semble important de créer des ressources évolutives pour qu'elles soient capables de rendre compte des phénomènes émergeant, plus ouvertes afin de faciliter leur mise à jour. Il semble aussi important de créer des ressources versionnées, afin d'intégrer au mieux les corrections d'erreurs trouvées au fil des travaux.

Nous proposons ici un corpus libre et évolutif annoté en entités nommées [1] afin de remédier au mieux aux problèmes cités plus haut. Ce corpus contient des textes très récents. Nous nous plaçons donc ici dans la continuité des travaux effectués par Salmon-Alt *et al.* (2004); Hernandez & Boudin (2013), dont l'idée est de fournir des corpus annotés libres. Nous nous plaçons dans la continuation des travaux d' Hernandez & Boudin (2013) en décidant d'annoter des données issues de la partie française de Wikinews, et que l'annotation en entités nommées que nous proposons se veut enrichir celle en parties-du-discours déjà présente. Cependant, nous nous en distinguons de par le type d'annotations que nous effectuons, nous avons ici privilégié une annotation manuelle sur une partie plus petite. Nous souhaitons, à terme, que ce corpus puisse être utilisé afin d'entraîner des systèmes automatiques pour des tâches d'extraction d'information, qui impliquent souvent d'effectuer en amont la reconnaissance des entités nommées, l'*entity linking* ainsi que l'extraction des relations qui lient les entités nommées.

1. disponible à l'adresse : `https://github.com/YoannDupont/WiNER-fr`

## 2 Le corpus et son annotation

Le corpus contient actuellement les contenus textuels des articles de Wikinews français des années 2016 à 2018, pour un total de 1191 articles. Nous avons pris l'ensemble des articles Wikinews, à l'exception d'un certain type de document : les tableaux de résultats sportifs. Ces documents contiennent typiquement une unique phrase décrivant la compétition ayant eu lieu suivi d'un tableau de scores.

Un unique annotateur ayant de fortes connaissances dans la reconnaissance des entités nommées a annoté l'ensemble des documents. Afin d'accélérer le processus d'annotation, nous avons utilisé un outil spécifique codé en python avec la librairie Tkinter (Shipman, 2013). L'outil permet de sélectionner des empans de texte et de leur attribuer un type d'une simple touche de clavier, avec la possibilité de diffuser l'annotation à l'échelle du document. Afin d'accélérer encore le processus d'annotation, nous avons également régulièrement entraîné un système par apprentissage afin de donner une pré-annotation pour un document. Nous nous sommes inspirés des systèmes décrits par Dupont (2017), qui donnent à notre connaissance des performances état-de-l'art sur le FTB. Plus précisément, nous avons utilisé le modèle utilisant des CRF (Lafferty *et al.*, 2001). Nous avons fait ce choix en prenant en compte le rapport entre la correction et le temps de traitement. Bien que la correction du système utilisant les Bi-LSTM-CRF (Lample *et al.*, 2016) soit meilleure, la rapidité à l'entraînement et à l'annotation du système utilisant des CRF permettaient d'avoir un processus d'annotation des documents globalement plus rapide. Fort *et al.* (2009) indiquent que la pré-annotation *« introduit un biais en faveur de la correction des pré-annotations, au détriment de la recherche de nouvelles EN »*. Pour réduire ce biais, nous avons procédé à l'annotation de chaque document pré-annoté en deux temps : un premier temps où les pré-annotations ont été corrigées et un second où le document était repris depuis le début à la recherche d'annotations manquées. Une fois le corpus annoté, nous avons utilisé un script pour détecter de potentielles incohérences. Le script crée d'abord un lexique par type d'entités en utilisant les différentes mentions annotés comme entrées. Le lexique ainsi créé est alors appliqué sur le corpus, les différences en type ou en frontière sont alors indiquées pour vérification par l'annotateur. Grâce à l'ensemble de ces méthodes afin d'accélérer le processus d'annotation, l'annotation manuelle, c'est-à-dire sans compter les temps d'entraînement du modèle par apprentissage, a pris une semaine à l'annotateur. Une révision manuelle de l'ensemble des annotations du corpus n'a pas encore été effectuée.

Nous avons annoté le corpus en nous basant sur un jeu simplifié des étiquettes du Quaero. Nous avons annoté les types suivants : les dates, les événements, les heures, les lieux, les organisations, les personnes et les produits (sans hiérarchie). Notre ensemble d'étiquette est globalement un sous-ensemble du jeu d'étiquettes défini par Quaero. Ce jeu d'étiquettes est comparable à celui de CoNLL-2003, où des entités temporelles et les événements remplacent le type "MISC". Les dates sont ici des dates dites absolues, c'est-à-dire que nous annotons uniquement les dates référant à un jour unique (1er janvier 2019), référant à un jour ou un mois identifiable sur un calendrier (1er janvier, janvier, janvier 2019), les décennies, les siècles, millénaires et les périodes clairement identifiées (par exemple, « les trente glorieuses » désignent une période précise). Nous n'annotons cependant pas les dates relatives comme « la veille », « le mois prochain ». Dans le cas d'une date relative par rapport à une date absolue (selon la définition précédente) comme « 1er janvier prochain », il a été décidé d'annoter « 1er janvier » comme une date absolue. La notion d'événements dans notre corpus est identique à celle de Quaero. Elle désigne, entre autres, les tournois sportifs (championnats du monde de patinage), les congrès, les événements annuels, les fêtes. Nous avons décidé d'annoter les événements climatiques ainsi que les affaires politico-juridiques, ces derniers étant laissés à l'appréciation de l'annotateur.

Nous avons donc ici décidé d'être le plus couvrant possible, ce qui permet de simplifier le travail d'annotation en réduisant les incertitudes par rapport à certains types d'événements. Les heures sont, à l'instar des dates, les heures dites absolues. Les lieux ici sont équivalents à l'ensemble des types de lieux définis dans le Quaero. Les organisations correspondent ici aux types « organisation » et « entreprise ». Les personnes désignent les personnes aussi bien réelles que fictives. Nous n'annotons que les prénoms, noms et surnoms. Les titres, fonctions, nationalités, etc. ne sont pas annotés. Ne sont pas annotés non plus les groupes de personnes, comme les familles lorsqu'elles sont désignées par leur nom de famille. Les produits désignent les différents produits qui sont présents dans le jeu d'étiquettes Quaero. Parmi eux, nous trouvons principalement les objets physiques (Airbus A380), les logiciels (jeux vidéos, etc.) et les produits médiatiques (émission de radio, TV, etc.).

Nous avons décidé d'annoter les composants des entités si ces derniers étaient également des entités nommées dans l'absolu. Par exemple, « Tour de France 2016 » est une entité de type événement contenant deux composants, à savoir « France » qui est annoté comme un lieu et « 2016 » qui est annoté comme une date. Nous avons décidé de le faire de manière systématique, même lorsque le lien n'est pas forcément évident. Par exemple dans « Université Leland Stanford Junior », « Leland Stanford Junior » réfère au fils des fondateurs, nous avons donc décidé de l'annoter. Un intérêt à cette annotation, malgré le biais positif qu'elle donne aux systèmes automatiques, est de faciliter le futur passage au schéma d'annotation Quaero. Des exemples d'annotations sont fournis dans la figure 1.

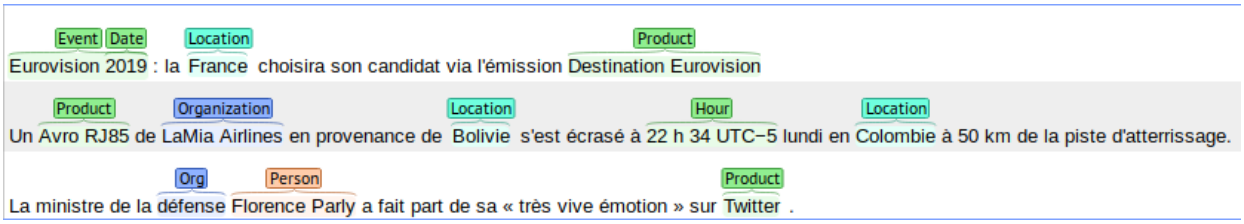



FIGURE 1 – Quelques exemples d'annotations visualisés avec l'outil BRAT.

Le corpus est en accès libre et se présente sous la forme d'un projet git hébergé sur GitHub. Nous utilisons un versionnement sémantique, c'est-à-dire un schéma "majeur.mineur.correctif". Une version majeure représente une annotation du corpus revue manuellement dans son intégralité. Une version mineure, l'ajout de nouveaux documents annotés avec des annotations non revues manuellement en vue d'une prochaine version majeure. Une version corrective implique uniquement la correction manuelle de l'annotation du corpus, aucun document ne pourra être ajouté.

La mise en ligne d'une nouvelle version majeure du corpus implique que ce dernier soit révisé. Si des portions du corpus ont déjà été vérifiées pour une version précédente du corpus et n'ont pas été modifiées depuis, leur vérification est optionnelle. Une version majeure du corpus ne pourra plus se voir attribuer de nouveaux documents, mais pourra toujours recevoir des *patchs* afin d'améliorer la correction des annotations. Des versions béta du corpus peuvent être diffusées de manière régulière selon l'avancement des annotations. À l'heure actuelle, nous souhaitons que ces versions béta correspondent à l'ajout d'années "pleines" (mois ou années) afin de limiter le nombre de sorties. Ces versions béta peuvent contenir des annotations non révisées.

Les ajouts et le corrections se font sur la base du volontariat. Afin de gérer au mieux les conflits, une fois la première version corrigée du corpus sortie, nous maintiendrons des branches séparées de la branche principale afin d'intégrer les modifications. Les mainteneurs du projet pourrons intégrer leurs changements simplement. Les participations d'une personne extérieure devront être faite via une

| type entité | compte | compte uniques |
|---|---|---|
| Date | 3817 | 1386 |
| Event | 720 | 340 |
| Hour | 945 | 426 |
| Location | 8719 | 2187 |
| Organization | 4215 | 1628 |
| Person | 5055 | 2499 |
| Product | 673 | 207 |
| global | 24144 | 8673 |

TABLE 1 – les comptes des entités par type. « uniques » compte les formes de surface différentes.

requête d'audit (*pull request*) et devront être vérifiées par un mainteneur du projet. Ces vérifications devront s'assurer de la cohérence des annotations avec le schéma.

# 3 Mesures relatives au corpus

Le corpus des années 2016 à 2018 de Wikinews français comporte un total de 1191 documents pour 322931 tokens (selon une segmentation automatique). Comme indiqué dans le tableau 1, le corpus comprend un total de 24144 annotations (pour 1122 annotations imbriquées, soit environ 4% du volume total des annotations).

Le corpus est actuellement annoté au format BRAT (Stenetorp *et al.*, 2012) : un premier fichier contient le contenu textuel du document et un second fichier contient les annotations déportées. Ce choix a deux motivations principales. Premièrement, il permet d'obtenir un historique très clair des modifications faites aux annotations dans un logiciel de gestion de version. Le second intérêt est que ce format est que nous pouvons utiliser directement l'outil BRAT afin de lier les mentions à une base (désambiguïsation des entités, ou *named entity linking*) ainsi que les relations qui lient les différentes entités entre elles. Ce choix est donc également motivé par les ajouts prévus au corpus.

Comme indiqué dans la section 2, nous avons utilisé un script afin de trouver des erreurs d'annotation de manière semi-automatique. Afin d'évaluer l'apport de l'utilisation de ce script, nous avons comparé la première et la dernière version de chaque fichier. En évaluant les différences, nous pouvons donner une estimation de l'apport d'une recherche d'erreurs simple. Nous avons calculé cet accord selon deux points de repères : en prenant l'ensemble des documents et en prenant uniquement ceux ayant subi au moins une modification par l'annotateur. Au total, 183 documents ont subi des modifications depuis leur première version, soit 15%. La f-mesure entre la première et la dernière version de chaque article du corpus vaut 0.981 sur l'ensemble du corpus.

Afin de fournir un indice de la stabilité des annotations manuelles, nous avons calculé un accord intra-annotateur (Krippendorff, 2018). Nous avons sélectionné au hasard 1 document par mois, pour un total de 36 documents, en ignorant les documents ayant subi au moins une correction entre leur première et leur dernière version. Les documents ont été annotés sans pré-annotation. Nous avons calculé l'accord intra-annotateur à l'aide de la f-mesure plutôt que du $\kappa$ de Cohen (1960). La raison de ce choix réside dans la nature des annotations, dont le nombre n'est pas fixé à l'avance et dont l'estimation de l'accord attendu servant de base de comparaison est difficile, comme noté par Grouin

| type entité | précision | rappel | f-mesure |
|---|---|---|---|
| Date | 0,9444 | 0,9754 | 0.9597 |
| Event | 0,625 | 0,7353 | 0,6757 |
| Hour | 0,9 | 1,0 | 0,9474 |
| Location | 0,9308 | 0,9528 | 0,9416 |
| Organization | 0,7576 | 0,7895 | 0,7732 |
| Person | 0,9439 | 0,9343 | 0,9391 |
| global | 0,9097 | 0,9371 | 0,9232 |

TABLE 2 – L'accord intra-annotateur pour chaque type et en global.

*et al.* (2011); Dalianis (2018). Nous avons ainsi pu calculer diverses métriques, données dans le tableau 2. Bien que la qualité globale atteigne 0,9232 de f-mesure, nous notons une variabilité de qualité en fonction du type d'entité. Ainsi, les dates, heures, lieux et personnes ont de très bons résultats, alors que ceux des organisations et des événements sont moindres.

La différence la plus fréquente est celle d'annoter plus d'entités (bruit), cette différence se répartit équitablement entre les lieux, organisations et événements. Les lieux étant plus nombreux, nous pouvons considérer que cette erreur est plus fréquente pour les organisations et les événements. Pour les organisations, ces différences sont principalement sur des mentions nominales et les ministères (« régime syrien », « économie »), pour les événements ces différences concernent surtout les événements politiques ou climatiques (« primaire socialiste », « Camp Fire »). Les erreurs de type, frontière ou silence ont des volumes comparables. Les erreurs de frontière concernent principalement le déterminant « le » (« l'armée syrienne » → « armée syrienne »), ou la nature d'un lieu (« l'autoroute E40 » → « E40 »). Ces erreurs semblent être principalement présentes dans les documents de l'année 2016, première année à avoir été annotée. Ces erreurs pourraient donc être dues à l'absence de mise-au-point d'une mini-référence et/ou d'une phase de rodage (Fort, 2012).

# 4 Comparaison avec d'autres corpus

Dans cette section, nous comparerons le corpus que nous avons produit avec d'autres corpus similaires, à savoir le FTB, Quaero, le Free-FTB et le corpus MEANTIME (Minard *et al.*, 2016).

Le corpus a une taille comparable au FTB en nombre de caractères. Parmi les types en commun, la principale différence est que nous ne faisons pas la distinction entre organisation et entreprise, différence faite dans le FTB. Une autre différence est que le FTB distingue les personnages fictifs des personnes réelles, distinction non faite ici. Notre schéma d'annotation est plus couvrant que celui du FTB : par exemple le FTB ne couvre pas les lieux géologiques, hydrologiques ou astrologiques.

Le corpus a une taille équivalente à environ un quart du Quaero. Un comparatif avec Quaero a été fait en section 2. La plupart des annotations imbriquées dans le corpus que nous proposons peuvent être remplacées par un composant de type "name" pour correspondre au modèle Quaero. Les autres composants doivent être récupérés manuellement.

Le Free-FTB est un corpus de textes issu d'articles de la partie française de Wikinews et de transcriptions d'Europarl. Ce corpus utilise les articles des années 2005 à 2012 de Wikinews. À l'heure

actuelle, le corpus Free-FTB ne comprend que des informations de segmentation et de POS, basées sur le FTB. Les annotations POS du Free-FTB sont les sorties d'un système par apprentissage entraîné sur le FTB. Le corpus que nous proposons ici, bien qu'ayant un schéma d'annotation différent du FTB, pourrait très bien être utilisé pour enrichir le Free-FTB d'information d'entités nommées.

Le corpus MEANTIME propose également une annotation en entités nommées sur les textes de 120 articles Wikinews et ses traductions en espagnol, italien et néerlandais (pour un total de 480 documents). Chaque corpus fait approximativement un dixième du corpus que nous proposons ici et les langues autres que l'anglais sont des traductions. Le corpus MEANTIME a cependant une annotation bien plus riche : les documents sont notamment annotés en chaines de coréférence des entités (intra- et inter-documents) et les relations entre entités. Le jeu d'annotation en entités nommées de MEANTIME se base sur celui de CoNLL 2003, enrichi d'expressions temporelles, produits et événements. L'une des différences notables entre MEANTIME et notre corpus est que dans MEANTIME les groupes sont également annotés, comme les groupes de personnes (*John Howard + Ian MacDonald*, *500 guests*, etc.) et les groupes d'organisations (*loss-making business*, etc.).

# 5 Conclusion et perspectives

Nous proposons ici un corpus annoté en entités nommées, libre et disponible. Il est également versionné, les ajouts, erreurs et corrections sont facilement traçables, des corrections peuvent également être proposées simplement. Nous souhaitons à cet effet que ce corpus soit collaboratif, des révisions (modification du corpus, du guide d'annotation, etc.) peuvent être effectuées sous réserve de validation. Étant donnés les points précédents, et tant que Wikinews sera alimenté de nouveaux articles, nous voulons ce corpus évolutif où de nouveaux textes seront ajoutés pour le maintenir à jour.

Il est prévu dans un premier temps de passer en revue de manière approfondie les annotations effectuées sur les années 2016 à 2018. Comme nous l'avons vu, bien que la plupart des types d'entités ont un fort accord intra-annotateur, les événements et les organisations ont un accord comparativement faible et méritent donc une plus grande attention. Nous prévoyons d'intégrer de la désambiguïsation des entités nommées ainsi qu'annoter les relations entre entités nommées, chose qu'il est possible d'effectuer avec l'outil BRAT. Le corpus continuera à être alimenté de nouveaux articles Wikinews afin de le maintenir le plus à jour possible. Lorsque ce dernier aura une taille comparable au corpus Quaero, nous pourrons l'annoter avec le même schéma d'annotation. Des campagnes d'évaluation pourront être effectuées en utilisant ce corpus à l'avenir, tant sur la reconnaissance des entités nommées que sur leur désambiguïsation. Il est également prévu d'enrichir le Free-FTB à l'aide de ce corpus afin d'obtenir un corpus segmenté et annoté en POS et entités nommées libre, accessible et de large volume. Nous pourrons à cette occasion refaire les expériences présentées par Hernandez & Boudin (2013) afin d'estimer si nous avons un apport similaire dans le cas des entités nommées par rapport au POS. Une autre expérience à mener serait d'utiliser notre corpus comme ressource supplémentaire pour des systèmes par apprentissage dans le cadre de tâches existantes.

# Remerciements

Les travaux présentés ici bénéficient en partie du soutien financier du projet PARSEME-FR (ANR-14-CERA-0001).

# Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for French. In A. ABEILLÉ, Ed., *Treebanks*. Dordrecht : Kluwer.

BOISEN S., CRYSTAL M., SCHWARTZ R. M., STONE R. & WEISCHEDEL R. M. (2000). Annotating resources for information extraction. In *LREC*.

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

DALIANIS H. (2018). Evaluation metrics and evaluation. In *Clinical Text Mining*, p. 45–53. Springer.

DODDINGTON G. R., MITCHELL A., PRZYBOCKI M. A., RAMSHAW L. A., STRASSEL S. & WEISCHEDEL R. M. (2004). The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation. In *LREC*, volume 2, p. 1–4.

DUPONT Y. (2017). Exploration de traits pour la reconnaissance d'entités nommées du français par apprentissage automatique. In *24e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, p. 42–55.

EHRMANN M. (2008). *Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation*. PhD thesis, Paris 7.

FINKEL J., DINGARE S., NGUYEN H., NISSIM M., MANNING C. & SINCLAIR G. (2004). Exploiting context for biomedical entity recognition : From syntax to the web. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, p. 88–91 : Association for Computational Linguistics.

FORT K. (2012). *Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus*. PhD thesis, Université Paris-Nord-Paris XIII.

FORT K., EHRMANN M. & NAZARENKO A. (2009). Vers une méthodologie d'annotation des entités nommées en corpus ? In *Traitement Automatique des Langues Naturelles 2009*.

GALIBERT O., QUINTARD L., ROSSET S., ZWEIGENBAUM P., NÉDELLEC C., AUBIN S., GILLARD L., RAYSZ J.-P., POIS D., TANNIER X. *et al.* (2010). Named and specific entity detection in varied data : The quæro named entity baseline evaluation. In *LREC*.

GALLIANO S., GRAVIER G. & CHAUBARD L. (2009). The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Tenth Annual Conference of the International Speech Communication Association*.

GRISHMAN R. & SUNDHEIM B. (1996). Message Understanding Conference-6 : A Brief History. In *COLING*, volume 96, p. 466–471.

GROUIN C., ROSSET S., ZWEIGENBAUM P., FORT K., GALIBERT O. & QUINTARD L. (2011). Proposal for an extension of traditional named entities : From guidelines to evaluation, an overview. In *Proceedings of the 5th Linguistic Annotation Workshop*, p. 92–100 : Association for Computational Linguistics.

HAN S., KWON S., YU H. & LEE G. G. (2017). Answer ranking based on named entity types for question answering. In *Proceedings of the 11th International Conference on Ubiquitous Information Management and Communication*, p. 71–74 : ACM.

HERNANDEZ N. & BOUDIN F. (2013). Construction automatique d'un large corpus libre annoté morpho-syntaxiquement en français. In *Traitement Automatique des Langues Naturelles (TALN)*.

KIM J.-D., OHTA T., TATEISI Y. & TSUJII J. (2003). GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, **19**(suppl 1), i180–i182.

KRIPPENDORFF K. (2018). *Content analysis : An introduction to its methodology*. Sage publications.

LAFFERTY J., MCCALLUM A. & PEREIRA F. (2001). Conditional random Fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML 2001*, p. 282–289.

LAMPLE G., BALLESTEROS M., SUBRAMANIAN S., KAWAKAMI K. & DYER C. (2016). Neural architectures for named entity recognition. *arXiv preprint arXiv :1603.01360*.

MINARD A.-L., SPERANZA M., URIZAR R., ALTUNA B., VAN ERP M., SCHOEN A., VAN SON C. *et al.* (2016). Meantime, the newsreader multilingual event and time corpus.

NOUVEL D., ANTOINE J.-Y., FRIBURGER N. & MAUREL D. (2010). An analysis of the performances of the casen named entities recognition system in the ester2 evaluation campaign.

NOUVEL D., EHRMANN M. & ROSSET S. (2015). *Les entités nommées pour le traitement automatique des langues*. Rapport interne, ISTE editions.

ROSSET S., GROUIN C., FORT K., GALIBERT O., KAHN J. & ZWEIGENBAUM P. (2012). Structured named entities in two distinct press corpora : Contemporary broadcast news and old newspapers. In *Proceedings of the Sixth Linguistic Annotation Workshop*, p. 40–48 : Association for Computational Linguistics.

SAGOT B., RICHARD M. & STERN R. (2012). Annotation référentielle du corpus arboré de paris 7 en entités nommées. In *Traitement Automatique des Langues Naturelles (TALN)*, volume 2.

SALMON-ALT S., BICK E., ROMARY L. & PIERREL J.-M. (2004). La freebank : vers une base libre de corpus annotés. In *Traitement Automatique des Langues Naturelles-TALN'04*, p. 10–p.

SEKINE S. & NOBATA C. (2004). Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy. In *LREC*.

SHIPMAN J. W. (2013). Tkinter 8.4 reference : a gui for python.

STENETORP P., PYYSALO S., TOPIĆ G., OHTA T., ANANIADOU S. & TSUJII J. (2012). Brat : a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, p. 102–107 : Association for Computational Linguistics.

SURDEANU M. (2013). Overview of the TAC2013 Knowledge Base Population Evaluation : English Slot Filling and Temporal Slot Filling. In *Text Analysis Conference*.

TJONG KIM SANG E. F. & DE MEULDER F. (2003). Introduction to the conll-2003 shared task : Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, p. 142–147 : Association for Computational Linguistics.

# Une approche hybride pour la segmentation automatique de documents juridiques

Filipo Studzinski Perotto[1] Fadila Taleb[2] Eric Trupin[1] Youssouf Saidali[1]
Maryvonne Holzem[2] Jacques Labiche[2] Laurent Vercouter[1]
(1) Normandie Université, UNIROUEN, UNIHAVRE, INSA Rouen, LITIS, 76000 Rouen, France
(2) Normandie Université, UNIROUEN, DYLIS, 76000 Rouen, France
filipo.perotto@litislab.fr

RÉSUMÉ

Cet article [1] propose une approche hybride pour la segmentation de documents basée sur l'agrégation de différentes solutions. Divers algorithmes de segmentation peuvent être utilisés dans le système, ce qui permet la combinaison de stratégies multiples (spécifiques au domaine, supervisées et non-supervisées). Un ensemble de documents étiquetés, segmentés au préalable et représentatif du domaine ciblé, doit être fourni pour être utilisé comme ensemble d'entraînement pour l'apprentissage des méthodes supervisées, et aussi comme ensemble de test pour l'évaluation de la performance de chaque méthode, ce qui déterminera leur poids lors de la phase d'agrégation. L'approche proposée présente de bonnes performances dans un scénario expérimental issu d'un corpus extrait du domaine juridique.

ABSTRACT

**A hybrid approach for automatic text segmentation**

This paper proposes a hybrid architecture for segmenting text documents, based on the aggregation of different solutions. Diverse segmentation algorithms can be incorporated into the system, allowing the combination of multiple strategies (domain-specific, supervised and unsupervised). A set of annotated documents is used for training the supervised methods, and for evaluating all the methods. The accuracy of each method determines its weight in the aggregation phase. A corpus extracted from a juridic domain was used for testing. The proposed approach presented good performances in such experimental scenario.

MOTS-CLÉS : segmentation linéaire automatique de texte.

KEYWORDS: automatic linear text segmentation.

## 1 Introduction

La *segmentation linéaire de texte* consiste à diviser un document en parties sémantiquement cohérentes, en identifiant des segments contigus et distincts en fonction de leurs caractéristiques communes (telle que la cohésion lexicale). L'automatisation de cette tâche est une étape aussi cruciale que problématique et les questionnements qu'elle soulève occupent une place importante au sein du *traitement automatique du langage naturel*. En effet, la segmentation constitue une porte d'entrée

1. Cet article est un résultat du projet *PlaIR 2.018*, cofinancé par l'Union Européenne à travers le Fonds Européen de Développement Régional (FEDER) et par la Région Normandie.

pour aborder d'autres problèmes, tels que la *fouille de textes*, la *synthèse de documents*, la *classification de documents*, la *recherche documentaire*, ainsi que la *visualisation*. La segmentation automatique est aussi l'étape préliminaire pour tout système d'aide à l'interprétation qui aurait pour but de faciliter l'accès à des documents complexes pour des lecteurs novices.

Tel est le cas du langage juridique utilisé dans les décisions de justice, qui en plus de la technicité et de l'arduité de leur vocabulaire, présentent une organisation textuelle particulière. Dans cet article, nous proposons une approche hybride pour la segmentation de documents issus d'un même corpus thématique. Plus spécifiquement, nous proposons un système dans lequel plusieurs algorithmes de segmentation peuvent être combinés automatiquement. Cela permet de faire collaborer différentes stratégies : des heuristiques fournies par un spécialiste du domaine, des mécanismes basés sur un apprentissage supervisé, et des méthodes non-supervisées. Pour pouvoir être intégré au système, un algorithme de segmentation doit être capable, lui étant donné un nouveau document à découper, de proposer un score pour chaque paragraphe en tant que candidat pour marquer le début d'un segment. L'architecture étant adaptative, des nouvelles méthodes peuvent être ajoutées au système à tout moment.

Un ensemble de documents préalablement segmentés (étiquetés avec les frontières correctes où un segment se termine et un autre commence) doit être fourni au système. Ces documents doivent être représentatifs du domaine ciblé. Ils servent comme exemples pour toutes les méthodes d'apprentissage supervisé, mais aussi comme ensemble de test permettant d'évaluer la précision de chaque méthode dans la tâche de segmentation. Cette évaluation de performance peut être interprétée comme une mesure de confiance sur chaque méthode, ce qui permet au système de leur attribuer un poids lors de l'agrégation des différentes solutions présentées dans une solution commune.

Cette approche a été testée sur un corpus extrait d'une base documentaire en ligne spécialisée dans la jurisprudence française en matière de transport. Nous avons segmenté un sous-ensemble de ces documents suivant l'usage rhétorique du domaine. Nous avons également défini un ensemble d'heuristiques, constituant un modèle linguistique qui est mis en œuvre par l'une des méthodes dans le système. L'approche hybride a permis d'identifier les segments corrects au sein des nouveaux documents avec une précision supérieure à celle de chaque méthode prise isolément.

Dans la suite de l'article, la section 2 réalise un bref aperçu des approches et des méthodes les plus importantes en segmentation automatique du texte. La section 3 présente le corpus de documents et la spécificité de la segmentation en question. La section 4 définie l'approche hybride que nous proposons. La section 5 compare expérimentalement notre architecture contre d'autres algorithmes classiques. Les conclusions et les travaux futurs sont discutés dans la section 6.

# 2 Travaux apparentés

La stratégie prédominante pour automatiser la segmentation linéaire de texte est l'utilisation des méthodes non-supervisées. Ces méthodes s'appuient sur la quantification de la cohésion lexicale entre différentes parties du document analysé (Koshorek *et al.*, 2018). La cohésion lexicale correspond à la manière dont les mots sont enchaînés dans le flux des phrases afin de créer des unités sémantiques (Morris & Hirst, 1991). Ces méthodes essaient d'identifier la cohésion lexicale dans une zone du texte, et ensuite de partitionner le document en un ensemble de segments thématiquement cohérents (Dadachev *et al.*, 2014). Les zones de texte avec un vocabulaire similaire sont susceptibles de faire

partie d'un même segment, et une variation lexicale peut être l'indicateur d'un changement de sujet.

`TextTiling` (Hearst, 1997) est le premier algorithme représentatif de cette approche. Il compare deux blocs de mots adjacents de longueur fixe en mesurant la répétition des mots entre eux. En déplaçant petit-à-petit les deux blocs tout le long du document, une fonction de similarité peut être dessinée selon la position (i.e. la frontière entre deux segments supposés). Les frontières qui affichent la plus petite similarité entre les blocs de texte qu'elles divisent sont sélectionnées comme candidates potentielles pour diviser le document en segments. Des méthodes plus performantes ont été proposées dans cette même approche grâce à l'utilisation des modèles statistiques plus fins (Choi, 2000; Brants *et al.*, 2002; Eisenstein & Barzilay, 2008; Chen *et al.*, 2009; Sakahara *et al.*, 2014), grâce à l'utilisation des relations sémantiques (ontologies) pour considérer la similarité entre les mots (Bayomi *et al.*, 2015; Ercan & Cicekli, 2016), ou des différentes façons d'extraire les caractéristiques du document (Utiyama & Isahara, 2001; Malioutov & Barzilay, 2006; Misra *et al.*, 2009; Dadachev *et al.*, 2014), ou par l'utilisation d'une représentation plus élaborée du lexique, s'éloignant du « bag-of-words » pour privilégier les « word embeddings » (Riedl & Biemann, 2012; Glavaš *et al.*, 2016).

Cependant, lorsqu'un ensemble représentatif de documents segmentés est disponible, l'utilisation de techniques d'apprentissage supervisé devient une opportunité intéressante. Pourtant peu de travaux suivant cette approche ont été publiés. (Beeferman *et al.*, 1999) utilise un algorithme de classification afin d'apprendre au système à détecter si une phrase donnée indique potentiellement le début ou la fin d'un segment. (Koshorek *et al.*, 2018) propose un modèle neuronal hiérarchique pour classifier l'appartenance d'une phrase à un segment spécifique. L'avantage d'utiliser l'apprentissage supervisé est que le critère de segmentation, même s'il est très complexe, est défini par extension. Si le jeu de données d'apprentissage est représentatif, il peut être « appris » au travers d'exemples, sans qu'une définition explicite soit nécessaire (Passonneau & Litman, 1997; Beeferman *et al.*, 1999).

# 3 Verrou scientifique

Dans le cadre d'un projet pluridisciplinaire [2], des chercheurs informaticiens et linguistes se sont posées la question de concevoir un système d'aide à l'interprétation d'un fond jurisprudentiel [3]. En amont de l'implémentation d'un tel système, un travail linguistique a été mené sur un corpus de plus de 300 décisions de justice dans le but de comprendre leur structure, le mécanisme argumentatif mis en œuvre, et surtout de mettre au jour des scénarios modaux [4] susceptibles de déclencher des parcours interprétatifs pouvant aider à la lecture de ces décisions (Taleb & Holzem, 2018). Dans cette perspective, une segmentation semi-manuelle a été effectué sur le corpus. Le découpage a servi à une première analyse textométrique différentielle [5].

Nous avons travaillé sur un corpus extrait d'une base documentaire en ligne de l'Institut du Droit

2. PlaIR (Plateforme d'Indexation Régionale - Normandie)

3. Parler d'aide à l'interprétation suppose d'adapter celle-ci aux pratiques professionnelles concernées et donc aux textes. L'enjeu est de pouvoir accéder finement à l'argumentaire juridique au sein de chacun des segments et comprendre leurs enchaînements intra et inter segmentales.

4. Scénario modal entendu comme l'expression du point de vue adopté par le magistrat en fonction des faits (modalité aléthique), puis de leur appréciation (modalité appréciative), autorisant un jugement de valeur de nature légal sur les actes en question (modalité axiologique), et le verdict (modalité déontique).

5. Repérage de constructions lexicales recourantes, qui marquent des moments clés du jugement, souvent corrélés à une transformation modale.

International des Transports (IDIT) [6]. Leur bibliothèque numérique, spécialisée dans la jurisprudence en matière de transport, dispose d'environ 40000 documents, dont 3000 en accès libre. Nous avons analysé un sous-ensemble de 300 arrêts [7], écrits en français, produits à différents moments, par différents juges, dans diverses *cours d'appel* [8] françaises. Chacun de ces documents a été manuellement partitionné en 4 segments, suivant l'usage en pratique dans le domaine. Ce corpus annoté constitue un ensemble de données d'entraînement permettant l'application de techniques d'apprentissage supervisé.

En raison de la spécificité du domaine, la structure des segments dans les documents du corpus est régulière. Chaque document présente $4$ segments distincts et contigus, apparaissant toujours dans le même ordre : (1) le *header*, où sont déclarées des informations telles que le nom du tribunal, le juge, la ville, la date et le nom des parties en litige ; (2) les *faits*, où le contexte du désaccord est rappelé sous la forme de récit, ainsi que les prétentions des parties appelantes et intimées, et le verdict prononcé par la juridiction d'instance inférieure (le tribunal) ; (3) les *motifs*, où le juge expose les arguments qui justifient sa décision ; et (4) la *conclusion*, où le juge énonce son verdict final. Ces segments constituent des unités de plusieurs paragraphes à l'intérieur du document. Les frontières entre les segments se trouvent toujours dans le passage d'un paragraphe au suivant [9]. Chaque paragraphe appartient nécessairement à un seul segment.

Formellement, le corpus est constitué d'un ensemble de $n$ documents $D = \{d_1, d_2, ..., d_i, ..., d_n\}$. Chaque document $d_i$ est une séquence de $m_i$ paragraphes $\{p_1, p_2, ..., p_x, ..., p_{m_i}\}$. La tâche consiste à rechercher les limites $\{b_1, b_2, b_3\}$ qui séparent correctement les $4$ segments cherchés $\{s_1, s_2, s_3, s_4\}$ (*header*, *faits*, *motifs*, *conclusion*), où $b_j$ correspond à l'index du paragraphe qui marque le début du segment $j + 1$. Chaque segment doit être constitué d'au moins un paragraphe. Le premier segment commence toujours au premier paragraphe, et le dernier segment se termine toujours par le dernier paragraphe. On observe donc la contrainte $1 < b_1 < b_2 < b_3 \leq m_i$. Le défi est celui de pouvoir segmenter des nouveaux documents, jamais vus auparavant, uniquement à partir de l'ensemble de documents d'entraînement.

# 4 Méthode Proposée

La première approche que nous avons adoptée a été l'analyse textométrique du sous-ensemble de 300 documents étiquetés avec les frontières réelles entre les segments, pour en déduire un système d'expressions régulières capables de capturer la majorité des cas (table 1). Même si ces documents juridiques issues des la cour d'appel française ne sont pas des formulaires mais des textes libres, il est fréquent d'observer l'usage des marqueurs explicites pour indiquer le début de chaque segment. Ces règles grammaticales permettent de retrouver correctement environs $90\%$ des frontières entre les segments dans le jeu de documents en question. C'est une performance assez correcte, mais qui signifie plutôt que dans la majorité des documents il y a une intention explicite de bien démarquer le passage entre les segments. En plus, on peut s'attendre à ce que le taux de précision baisse dès lors que les règles seront appliquées à d'autres documents, en dehors du sous-ensemble qui leur a donné

6. `www.idit.fr`

7. Le contenu des documents est disponible en format texte brut, généralement extrait des fichiers PDF. Après prétraitement, chaque document est représenté comme une collection ordonnée de paragraphes.

8. Cette juridiction de second degré est sollicitée par l'une des parties d'un litige ayant fait appel d'un précédent jugement rendue par une juridiction de premier degré.

9. Un changement de segment est contraint de coïncider avec un changement de paragraphe.

| Segment | Expression régulière (insensible à la casse) | Confiance |
|---|---|---|
| Faits | ^(\s*)(*exposé* (*du litige* \| *des faits*) \| *faits*)([\s :]*)$ | 1.0 |
| Faits | ^(\s*)(*vu que* \| *par actes* \| *la cour* (.*) *attendu que* ) | 0.9 |
| Motifs | ^(\s*)(*motifs*)([\s :]*)$ | 1.0 |
| Motifs | ^(\s*)(*sur ce* \| (*ce*) ? *sur quoi* \| (*ceci* \| *cela*) *étant* \| *étant exposé*) | 0.9 |
| Conclusion | ^(\s*)(*décision*)([\s :]*)$ | 1.0 |
| Conclusion | ^(\s*)(*par ces motifs*) | 0.9 |

TABLE 1 – Expressions régulières construites après analyse textométrique pour identifier les expressions-repères indiquant le début de chaque segment dans les documents de décision de cours d'appel françaises.

origine. Finalement, la faiblesse de cette méthode vient des $10\%$ de segmentations incorrectes. Ces sont les cas pour lesquels aucune règle ne peut être appliquée, et donc aucune suggestion ne peut être retournée, ou les cas où plusieurs correspondances sont trouvées.

Le défi posé par ce problème est donc de développer une méthode plus précise que celle basée sur les heuristiques apportées par un expert humain. Les méthodes non-supervisées, comme on peut le constater dans les résultats (section 5), ont une précision assez faible (inférieure à $10\%$). Ce n'est pas surprenant, vu qu'il s'agit de documents longs, segmentés d'une manière spécifique au domaine, assez éloignée du principe qui régit la segmentation thématique, qui consiste à trouver des segments homogènes d'un point de vue sémantique. Le corpus en question demande plutôt une tâche de structuration, chaque document étant composé d'un nombre identique de segments, qui suivent une structure canonique bien précise. Cela laisse supposer que l'on retrouve un vocabulaire commun tout au long du document dans ce type de décision juridique, alors que quelques mots structurants (repérés par l'approche textométrique) viennent marquer les changement de segment, de sorte que les approches basées sur la cohésion lexicale ne peuvent pas être très performantes.

Finalement, les méthodes basées sur l'apprentissage supervisé s'approchent de la précision rendue par la méthode d'expressions régulières, restant pourtant moins performantes. Dans cet article, l'apprentissage supervisée a été réalisé de la façon suivante : un classificateur du type *Naive Bayes* a été entraînée pour identifier les paragraphes initiaux des différents segments. Les exemples d'entraînement étant constitués de la liste de *tokens* (mots après filtrage de *stop-words*) contenus dans le paragraphe, et l'étiquette du segment qu'il débute ($1 =$ *entête*, $2 =$ *faits*, $3 =$ *motifs*, $4 =$ *conclusions*, ou $0$ s'il n'est au début d'aucun segment). Cette méthode a pu atteindre un taux de précision d'environs $80\%$.

La solution que nous proposons consiste à faire collaborer plusieurs méthodes de segmentation dans un système hybride. Lorsque nous disposons d'un ensemble représentatif d'exemples indiquant comment les documents à l'intérieur du corpus doivent être segmentés, il est possible d'estimer la précision de chaque méthode. Cette valeur (la précision de chaque méthode évaluée contre le jeu de documents préalablement segmentés) indique la confiance que le système a en la méthode. Cette confiance servira ensuite à pondérer ses réponses lors d'une phase ultérieure d'agrégation de la solution. Il s'agit d'un problème d'*agrégation de préférences* (Conitzer, 2006; Sen & Yang, 1994). Il faut combiner de manière adéquate les différents résultats provenant des différentes méthodes. Dans le cas de la segmentation, une moyenne pondérée de différentes positions proposées ne semblerait pas avoir beaucoup de sens. Nous proposons de suivre la majorité pondérée par deux facteurs : la confiance qui le système a dans la méthode, et la confiance que la méthode a en ses réponses.

Formellement, étant donné un ensemble de documents $D = \{d_1, d_2, ..., d_i, ..., d_n\}$, où les frontières

$B_i = \{b_1, b_2, b_3\}$ entre les segments sont connues pour chaque document $d_i$, et un ensemble de $z$ méthodes de segmentation $A = \{a_1, a_2, ..., a_k, ..., a_z\}$, nous calculons la qualité $w_k$ de la méthode $a_k$ dans la segmentation de l'ensemble de documents. Les métriques classiques de précision, telles que celles proposées par (Beeferman *et al.*, 1999) et (Hearst, 1997), considèrent la distance de chaque phrase par rapport à son segment correct. Nous avons choisi un score plus simple et plus strict. La précision $w_k$ d'une méthode $a_k$ correspond à la proportion de frontières correctement trouvées, et est calculée comme suit :

$$w_k = \frac{\sum_{i=1}^{n} \sum_{j=1}^{3} c_{k,i,j}}{3n} \tag{1}$$

où $n$ est le nombre de documents dans $D$, 3 est le nombre de frontières à trouver dans chaque document, $c_{k,i,j} = 1$ si la frontière $b_j$ dans le document $d_i$ est correctement proposée par la méthode $a_k$, et 0 sinon. Cette formule peut être utilisée telle quelle pour évaluer les méthodes non-supervisées, ou basées sur des heuristiques. Pour les méthodes supervisées, considérant que l'ensemble de test est utilisé comme ensemble d'entraînement, on utilise une validation croisée du type 5-*fold* [10].

Après avoir analysé un document donné $d_i$, une méthode $a_k$ doit indiquer une valeur de confiance $v_{k,i,j,x}$ indiquant sa croyance que la frontière $b_j$ se trouve au début de chaque paragraphe $p_x$ du document. Une confiance combinée peut être en suite calculée à partir des valeurs de confiance de chaque méthode individuelle, comme suit :

$$\forall i \; \forall j \; \forall x \quad : \quad s_{i,j,x} = \sum_{k=1}^{z} {w_k}^2 \; v_{k,i,j,x} \tag{2}$$

où $s_{i,j,x}$ est le score du paragraphe $p_x$ en tant que candidat pour la frontière $b_j$ du document $d_i$, $w_k$ est la confiance accordée à la méthode $a_k$, et $v_{k,i,j,x}$ est la confiance de la méthode $a_k$ sur le fait que la frontière $b_j$ du document $d_i$ se trouve au paragraphe $p_x$. Dans la formule, on utilise le carré de la confiance pour favoriser les réponses données par les meilleures méthodes. La segmentation finale d'un document donné $d_i$ est choisie en sélectionnant la combinaison de frontières qui obtient le plus grand score moyen, en respectant la contrainte d'ordre entre les segments, i.e. celle qui optimise :

$$\max \frac{1}{3} \sum_{j=1}^{3} s_{i,j,x} \quad \text{sujet à} \quad b_1 < b_2 < b_3 \tag{3}$$

# 5 Résultats

Nous avons essayé notre approche hybride en utilisant 3 différentes méthodes :

(1) La première méthode implémente l'ensemble d'expressions régulières issues d'une analyse textométrique sur l'ensemble de documents d'entraînement (table 1). Ces règles sont conçues pour rechercher l'occurrence de certaines expressions-repères qui indiquent le début de chaque segment. Chacune de ces règles est associée à un niveau de fiabilité (aussi fourni par l'expert du domaine). Lorsque l'expression est trouvée dans le texte, cette fiabilité de la prédiction est retournée, associée

10. L'ensemble de documents d'entraînement est reparti en 5 différents échantillons. L'apprentissage en utilise 4 et la validation se fait sur l'échantillon restant. La démarche est répétée pour chaque échantillon.

au paragraphe en question, pour la frontière en question. Si aucune règle ne s'applique au paragraphe, la valeur 0 est retournée.

(2) La deuxième méthode est une version adaptée de l'algorithme non-supervisé classique `TextTiling` (Hearst, 1997). Comme les limites de segment sont contraintes de coïncider avec les limites de paragraphe, les comparaisons de similarité ne sont faites que sur ces positions. `TextTiling` retourne, pour chaque paragraphe, une valeur qui correspond à la probabilité que ce paragraphe soit le début d'un segment, sans pouvoir préciser de quel segment s'agit-il.

(3) La dernière méthode utilise un classificateur bayésien naïf, entraîné pour prévoir la probabilité qu'un paragraphe donné soit le paragraphe initial d'un segment cherché.

Les deux premières méthodes n'ont pas besoin d'être entraînés sur les documents préalablement étiquetés : la première est constituée d'heuristiques, et la deuxième est non-supervisée. Leur évaluation peut être faite directement en regardant leur précision dans la segmentation du jeu de documents d'entraînement, ici servant à tester leur performance. La dernière méthode utilise l'apprentissage supervisé. Dans ce cas, même si l'entraînement se fait avec l'intégralité de la base de documents étiquetés, sa précision doit être évaluée par validation croisée.

Dans la table 2, nous pouvons comparer la précision de chacune des méthodes séparément, puis la précision de notre système hybride, qui met en place une collaboration entre eux. Nous pouvons constater que l'approche hybride conduit à une amélioration de la performance, en comparaison avec chacune des autres méthodes isolées.

| **Stratégie** | **Méthode** | **Précision** |
|---|---|---|
| *heuristique* | `RegEx` | 0.91 |
| *non-supervisée* | `TextTilling` | 0.08 |
| *supervisée* | `NaiveBayes` | 0.81 |
| *combinée* | **`Hybride`** | **0.96** |

TABLE 2 – Comparaison de performance entre les méthodes expérimentées.

# 6 Conclusions et Travaux Futurs

Dans cet article, nous proposons une architecture hybride pour la segmentation linéaire de documents texte. Dans cette architecture, chaque méthode peut implémenter un algorithme différent, ce qui permet de combiner la puissance de plusieurs stratégies : spécifiques au domaine, supervisées et non-supervisées. Un ensemble de documents extraits du domaine juridique préalablement segmentés a été utilisé pour l'entraînement des méthodes supervisées et pour l'évaluation de toutes les méthodes. Nous avons pu démontrer que l'agrégation des différentes solutions à l'aide d'une règle de majorité pondérée a permis d'améliorer la précision de la segmentation automatique si comparé à la performance de chaque méthode isolée. Comme chaque méthode cherche à identifier des caractéristiques différentes pour déterminer les changements de segments, même une méthode qui n'est pas très performante en moyenne peut venir contribuer à la solution combinée dans les cas où les méthodes plus performantes échouent. Nous souhaitons poursuivre cette recherche en réalisant une analyse comparative plus approfondie entre notre méthode et d'autres algorithmes de segmentation, ainsi qu'en étendant les tests à d'autres jeux de documents.

# Références

BAYOMI M., LEVACHER K., GHORAB M. R. & LAWLESS S. (2015). Ontoseg : A novel approach to text segmentation using ontological similarity. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, p. 1274–1283 : IEEE.

BEEFERMAN D., BERGER A. L. & LAFFERTY J. D. (1999). Statistical models for text segmentation. *Machine Learning*, **34**(1-3), 177–210.

BRANTS T., CHEN F. & TSOCHANTARIDIS I. (2002). Topic-based document segmentation with probabilistic latent semantic analysis. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, CIKM '02, p. 211–218, New York, NY, USA : ACM.

CHEN H., BRANAVAN S. R. K., BARZILAY R. & KARGER D. R. (2009). Global models of document structure using latent permutations. In *Proceedings of Human Language Technologies : The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, p. 371–379, Stroudsburg, PA, USA : Association for Computational Linguistics.

CHOI F. Y. Y. (2000). Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL 2000, p. 26–33, Stroudsburg, PA, USA : Association for Computational Linguistics.

CONITZER V. (2006). *Computational Aspects of Preference Aggregation*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, USA. AAI3232648.

DADACHEV B., BALINSKY A. & BALINSKY H. (2014). On automatic text segmentation. In *Proceedings of the 2014 ACM Symposium on Document Engineering*, DocEng '14, p. 73–80, New York, NY, USA : ACM.

EISENSTEIN J. & BARZILAY R. (2008). Bayesian unsupervised topic segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, p. 334–343, Stroudsburg, PA, USA : Association for Computational Linguistics.

ERCAN G. & CICEKLI I. (2016). Topic segmentation using word-level semantic relatedness functions. *J. Inf. Sci.*, **42**(5), 597–608.

GLAVAŠ G., NANNI F. & PONZETTO S. P. (2016). Unsupervised text segmentation using semantic relatedness graphs. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, p. 125–130 : Association for Computational Linguistics.

HEARST M. A. (1997). Texttiling : Segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, **23**(1), 33–64.

KOSHOREK O., COHEN A., MOR N., ROTMAN M. & BERANT J. (2018). Text segmentation as a supervised learning task. *CoRR*, **abs/1803.09337**.

MALIOUTOV I. & BARZILAY R. (2006). Minimum cut model for spoken lecture segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, p. 25–32, Stroudsburg, PA, USA : Association for Computational Linguistics.

MISRA H., YVON F., JOSE J. M. & CAPPE O. (2009). Text segmentation via topic modeling : An analytical study. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, p. 1553–1556, New York, NY, USA : ACM.

MORRIS J. & HIRST G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Comput. Linguist.*, **17**(1), 21–48.

PASSONNEAU R. J. & LITMAN D. J. (1997). Discourse segmentation by human and automated means. *Computational Linguistics*, **23**(1), 103–139.

RIEDL M. & BIEMANN C. (2012). Topictiling : A text segmentation algorithm based on lda. In *Proceedings of ACL 2012 Student Research Workshop*, ACL '12, p. 37–42, Stroudsburg, PA, USA : Association for Computational Linguistics.

SAKAHARA M., OKADA S. & NITTA K. (2014). Domain-independent unsupervised text segmentation for data management. In *2014 IEEE International Conference on Data Mining Workshop*, p. 481–487 : IEEE.

SEN P. & YANG J.-B. (1994). Design decision making based upon multiple attribute evaluations and minimal preference information. *Mathl. Comput. Modelling*, **20**(3), 107–124.

TALEB F. & HOLZEM M. (2018). Exploration textométrique d'un corpus de motifs juridiques dans le droit international des transports. In *Proceedings of the* $14^{th}$ *Int. Conf. on Statistical Analysis of Textual Data (JADT 2018)*.

UTIYAMA M. & ISAHARA H. (2001). A statistical model for domain-independent text segmentation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, p. 499–506, Stroudsburg, PA, USA : Association for Computational Linguistics.