Le traitement automatique des langues face à l'évolution des usages de la langue

Cyril Grouin

Habilitation à diriger des recherches

UNIVERSITÉ PARIS-SACLAY

Mémoire

présenté pour obtenir

l'Habilitation à Diriger des Recherches

Spécialité Informatique

ED 580 — Sciences et Technologies de l'Information et de la Communication

LISN — Laboratoire Interdisciplinaire des Sciences du Numérique

Le traitement automatique des langues face à l'évolution des usages de la langue

Par : Cyril GROUIN

Marraine scientifique :

Mme Anne VILNAT, Professeure des Universités en Informatique, Université Paris-Saclay

Rapporteurs :

Mme Béatrice DAILLE, Professeure des Universités en Informatique, Université de Nantes
M. Patrick RUCH, Professeur des Universités en informatique, Haute École Spécialisée de Suisse Occidentale (HES-SO), Genève
M. Mathieu VALETTE, Professeur des Universités en sciences du langage, Institut National des Langues et Civilisations Orientales (INaLCO), Paris

Examinatrices :

Mme Pascale SÉBILLOT, Professeure des Universités en informatique, Institut National des Sciences Appliquées (INSA), Rennes
Mme Anne VILNAT, Professeure des Universités en informatique, Université Paris-Saclay

Date de soutenance : 23 mars 2023

Table des matières

Chapitre

Remerciements

Le travail rassemblé et synthétisé dans ce manuscrit provient de collaborations diverses, avec plusieurs personnes que je souhaiterais remercier ici, tant pour le travail accompli que pour les discussions et échanges que nous avons pu avoir, quel qu'en fut le thème.

Parce qu'il s'agit d'une habilitation à diriger des recherches, je voudrais remercier en premier lieu les doctorants que j'ai co-encadrés : Liyun Yan, Alexandra Benamar, et Atilla Alkan Kaan. Vous aviez chacun une manière personnelle d'aborder votre thèse, reflet de vos parcours étudiants et de vos états d'esprit personnels. Ces trois thèses se sont réalisées dans des contextes et domaines très différents. Cette diversité m'a permis d'entrevoir les possibilités dans lesquelles se déroulent les thèses et travaux d'encadrement de recherches. Je tiens également à remercier les titulaires d'une HDR qui ont été les directeurs officiels de ces thèses et qui m'ont fait confiance en m'offrant l'opportunité d'encadrer ces thèses à leurs côtés : Mathieu Valette pour la thèse de Liyun, Anne Vilnat pour celle d'Alexandra, et enfin Fabian Schussler et Pierre Zweigenbaum pour la thèse d'Atilla, ainsi que les acteurs de collaborations industrielles, en particulier Meryl Bothua (EDF R&D).

Bien que ne relevant pas d'un encadrement à temps plein, ma participation aux comités de pilotage des thèses effectuées à Santé Publique France, notamment celle de Yasmine Baghdadi (co-dirigée par Anne Gallay et Anne Fouillet), a donné lieu à des collaborations fructueuses que je n'imaginais pas lors des premiers CoPil, et que j'ai réellement appréciées.

Je souhaite également remercier les post-doctorants avec qui j'ai eu le plaisir de travailler : Eva D'Hondt, François Morlane-Hondère, Leonardo Campillos-Llanos, et Lucie Gianola. Pour vous, il n'était plus question de vous encadrer, mais plutôt de vous guider ou de vous sortir d'une impasse méthodologique. Soyez assurés que les discussions que nous avons pu avoir ont été pour moi enrichissantes du fait de vos parcours respectifs.

Je n'oublie pas les stagiaires venus au LIMSI ou au LISN avec qui j'ai collaboré (Alix, Cyril, Dalia, Lufei, Philippe) et qui ont chacun apporté une contribution à ce travail de recherche, ainsi que les étudiants de l'INaLCO dont j'ai encadré le mémoire du master TAL (Aleksandra, Aurélie, Benjamin, Catherine, Clémence, Elise, Elvira, Emilie, Fatemeh, Geneviève, Giovanna, Haruka, Irina, Jia, Lara, Levana, Liyun, Lucía, Lufei, Maria, Marine, Mei, Morgane, Nidia, Qi, Selen, Sotiria, Svetlana, Véronique, Virginie, Xianfan, Yingying, Yunbei). Certains parmi vous ont pris goût à la recherche et ont poursuivi leurs efforts en thèse, ce dont je me réjouis.

Je remercie Anne d'avoir accepté de parrainer ce travail sur le plan scientifique, Béatrice, Mathieu, et Patrick pour avoir rapporté ce travail, et Pascale pour avoir accepté de faire partie du jury. Merci également à Anne et Gabriel pour la relecture du manuscrit et leurs commentaires.

Je profite de cette page pour remercier tous mes co-auteurs d'articles, beaucoup trop nombreux pour être listés, mais qui ont contribué à façonner mon travail de recherche, et dans certains cas à relativiser certains échecs. J'ai également une pensée pour les amis du laboratoire pour la richesse des échanges et la bonne ambiance propice à la réflexion. Je félicite Véronique Moriceau et Laure Soulier pour leur propre HDR. Nous avons souvent échangé et comparé les démarches administratives à accomplir pour réaliser une habilitation selon qu'on est inscrit à l'université Paul-Sabatier à Toulouse, à Sorbonne Université ou à Paris-Saclay.

Enfin, je remercie les collègues du laboratoire qui nous accompagnent au quotidien dans nos recherches mais bien souvent de manière invisible : administration, bibliothèque, gestion, logistique, service informatique. Non seulement votre travail n'est pas simple, mais en plus il est souvent oublié...

Chapitre 1

Introduction générale

Sommaire

1.1 Le traitement automatique des langues et les problèmes résolus

Ce manuscrit s'ouvre sur un constat régulièrement observé en traitement automatique des langues (TAL), et plus précisément dans ce qui est qualifié de « TAL robuste » par opposition au « TAL théorique » [Cori, 2008]. Quelle que soit la tâche considérée, et plus particulièrement pour les tâches de bas niveau (tokénisation, découpage en phrases, etc.) ou de base (étiquetage en parties du discours, repérage d'entités nommées, etc.), il est d'usage de considérer qu'un problème est résolu lorsque qu'un optimum est atteint par les méthodes ou systèmes élaborés pour ladite tâche. Cet optimum est mesuré par une évaluation quantitative au moyen de mesures faisant consensus dans la communauté scientifique concernée, telles que les mesures de rappel, précision et F-mesure [Paroubek, 2013]. Ce constat semble s'être renforcé ces dernières années sous l'effet de l'incitation à publier pour ne pas périr au niveau professionnel[1], engendrant de nombreuses publications scientifiques mettant chacune en avant des valeurs numériques d'évaluation forcément les plus élevées possible. Conclure que le problème auquel on s'attaquait est résolu me paraît précipité pour au moins deux raisons.

La première raison concerne l'objet d'étude en lui-même. Si le protocole expérimental est correctement défini et suivi, une évaluation est réalisée sur un jeu de données qui aura été réservé à cet effet pour mesurer les performances réelles du système développé, à l'image de ce qu'il se fait dans les campagnes d'évaluation traditionnelles (TREC, SemEval, etc.). Les propriétés présentes dans le jeu de test sont généralement similaires à celles observées dans le jeu d'entraînement. En conséquence, conclure qu'un problème est résolu ne me semble pouvoir être énoncé que si le système évalué est appliqué sur le plus grand nombre possible de corpus d'évaluation ou a minima représentatifs d'un maximum de propriétés à traiter, en obtenant les scores les plus élevés sur chacun de ces corpus, et si les annotations de ces corpus d'évaluation correspondent aux attendus de la communauté scientifique pour la tâche visée. Autrement dit, il faut rassembler plusieurs conditions pour pouvoir assurer qu'une tâche n'est plus un problème pour les systèmes du traitement automatique des langues. Cependant, cette première conclusion est remise en cause par une deuxième raison qui apparaît universelle.

1. Cette incitation est résumée en anglais par la formule *"Publish or Perish"*.

La deuxième raison concerne la langue, ses usages, et les usagers. Benveniste rappelle que « *C'est dans et par le langage que l'homme se constitue comme sujet* » [Benveniste, 1976, p. 259]. Cette observation rejoint le paradoxe saussurien : « *L'aspect social de la langue s'étudie sur n'importe quel individu, mais l'aspect individuel ne s'observe que dans le contexte social.* » [Labov, 1972]. L'apparition du web 2.0 et des micro-blogs en 2006 avec principalement Twitter et sa limitation initiale à 140 caractères, passée à 280 caractères en septembre 2016, a imposé aux utilisateurs de synthétiser leur pensée, mais cette limite les a également incités à utiliser des abréviations, des *emojis* (« émoticônes », voir p. 24), et à indexer leurs messages avec des mots-clés rassemblés sous des *hashtags* (« mot-dièse », voir p. 25). Même si des formulations syntaxiquement incorrectes et les abréviations existaient déjà dans les télégrammes (facturés au nombre de mots) ou dans les SMS (initialement limités à 160 caractères), leurs usages et la personnalisation de ces usages se sont amplifiés avec la démocratisation des SMS et l'avénement des réseaux sociaux. Rappelons que les langues évoluent sans cesse, par les évolutions scientifiques technologiques qui appellent de nouveaux concepts, d'abord obtenus par emprunt à d'autres langues (*code-switching* temporaire[2], voir p. 19) avant une francisation par créativité morphologique, et par les usages (réseaux sociaux, militantisme par le français inclusif, p. 27). Qualifier de résolue la plupart des tâches en traitement automatique des langues ressemble réellement à une chimère...

1.2 Du traitement automatique des langues à la science des données

Depuis le formalisme de la structure distributionnelle des langues mis en avant par Zellig Harris autour des régularités et relations distributionnelles entre éléments d'une langue, et notamment des relations entre phrases [Harris, 1954] d'une part, et les travaux de son élève Noam Chomsky sur les structures syntaxiques[3] et la grammaire générative [Chomsky, 1957] d'autre part, on assiste ces dernières années à une évolution rapide des techniques de représentation du texte fondées sur des approches distributionnelles. Contrairement aux travaux d'Harris qui portaient sur la sémantique distributionnelle de manière globale, les modèles de langue actuels n'encodent pas une sémantique globale, en tant qu'ensemble de connaissances sur le monde, comme celle qu'un locuteur peut constituer par expérience, mais uniquement la sémantique contenue dans les textes sur lesquels ils ont été entraînés [Sahlgren, 2008]. Cette approche rejoint la vision de la linguistique de corpus qui, « *sans renoncer à l'élaboration théorique, en limite la portée aux corpus étudiés, et sans se satisfaire de la seule démarche déductive, procède par essais et erreurs* » [Rastier, 2011]. En conséquence, et pour approximer cette sémantique globale, les approches distributionnelles s'appuient sur un grand volume de textes issus de sources variées, dans l'espoir que la sémantique contenue dans ces textes soit la plus diversifiée possible.

Dans ses travaux sur le fonctionnement de la langue, Ferdinand de Saussure distingue « *deux ordres de coordination* » entre termes [Saussure, 1916]. Les relations syntagmatiques relèvent de la syntaxe. Elles reposent sur la co-occurrence des mots dans une phrase, dans un paragraphe, ou dans un texte. Il s'agit de relations *in presentiæ* entre entités linguistiques. À l'opposé, les relations paradigmatiques ou associatives désignent les mots qui partagent des contextes mais ne cooccurrent jamais. On parle alors de relations qui unissent des termes *in abstentiæ*. On retrouve cette distinction dans les implémentations réalisées pour le traitement automatique des langues, avec les approches linéaires qui reposent sur des relations syntagmatiques, par opposition aux plongements qui s'appuient sur des relations paradigmatiques (voir section 5.1).

2. Par exemple, la marque « *Walkman* », créée en 1979 et entrée dans le *Larousse* en 1981, a servi à nommer temporairement l'équipement audio portatif. Il sera renommé « baladeur » en février 1983, sous l'impulsion de Georges Fillioud, Ministre de la Communication, suite aux travaux des commissions ministérielles de terminologie (https://www.lemonde.fr/archives/article/1983/02/17/cent-mots-pour-l-audiovisuel_2842859_1819218.html).

3. S'appuyant sur son fameux exemple d'une phrase correcte sur le plan syntaxique mais sans cohérence sur le plan sémantique *"Colorless green ideas sleep furiously"* (« D'incolores idées vertes dorment furieusement »).

La description linguistique englobe les ordres syntagmatique et pragmatique précédemment présentés, ainsi qu'un ordre référentiel qui fait le lien entre concepts et phrases au moyen d'images mentales, et un ordre herméneutique. L'ordre herméneutique renvoie aux conditions de production des textes, en incluant les informations pragmatiques, pour permettre l'interprétation de ces textes, à l'image de l'interprétation des écritures saintes. Rastier rappelle que cet ordre est « *inséparable de la situation historique et culturelle de la production et de l'interprétation* » [Rastier, 1996]. Les informations pragmatiques, parce qu'elles renvoient à une dimension culturelle et sociétale, sont plus complexes à identifier, mais peuvent s'avérer utiles, notamment en fouille d'opinion (voir chapitre 3).

Une avancée majeure des modèles de langue actuels repose sur le fait qu'ils sont entraînés conjointement pour traiter plusieurs tâches, partant du principe que des caractéristiques utiles pour une tâche seront bénéfiques pour les autres tâches [Collobert and Watson, 2008]. Cette approche s'oppose aux classifieurs linéaires, entraînés séparément sur chaque tâche, avec le risque d'une propagation d'erreurs en cas d'enchaînement en cascade. Au niveau informatique, cet objectif nécessite une puissance de calcul importante, en lien avec les paramètres d'apprentissage choisis [4]. L'augmentation des puissances de calcul [5] et l'intérêt porté par de grands acteurs de l'internet (Allen AI, Facebook, Google, HuggingFace) contribuent à dynamiser ce champ de recherche. Néanmoins, ces modèles sont coûteux à produire, et leur utilisation n'est pas neutre non plus. Des solutions alternatives pour « verdir » ces modèles ont été envisagées (voir section 5.1.2 et une solution que nous avons étudiée en section 5.4).

Nous observons que cette rapidité est telle que les articles qui décrivent ces méthodes et modèles sont d'abord déposés sur la plateforme d'articles arXiv.org, tant comme preuve de dépôt que pour communiquer immédiatement, plutôt qu'en conférences ou en revues dont les processus de publication sont moins immédiats. Dans un second temps, les auteurs passent par un processus de soumission plus traditionnel pour une parution dans les actes d'une conférence ou d'un workshop.

Les travaux des *data scientists* autour de la science des données constituent une révolution pour le traitement automatique des langues, qui doit forcer les TAListes à s'en emparer pour y réinjecter des connaissances linguistiques.

1.3 Synthèse des contributions

1.3.1 Déterminer les niveaux et types d'informations pour l'annotation

Mes travaux en repérage d'entités nommées et fouille d'opinion reposent sur des annotations manuelles. J'ai exploré l'intérêt de varier le niveau de détail dans les annotations et leur impact sur les performances des approches par apprentissage statistique. En domaine de spécialité, j'ai montré que les schémas fondés sur des unités lexicales courtes et des relations d'expansion constituent une meilleure approche que de travailler sur des portions sémantiquement pertinentes mais plus complexes à capturer [Morlane-Hondère et al., 2016b]. J'ai également étudié la possibilité de passer d'un schéma à un autre en variant le niveau de détails des annotations [Grouin, 2018]. En fouille d'opinion, j'ai observé que les inférences capturent des informations culturelles et sociétales qui sont complémentaires au repérage des mots porteurs d'émotion–sentiment–opinion pour déterminer la valence d'une expression dans des avis en mandarin standard [Yan, 2018, Yan et al., 2020].

4. Les ressources informatiques et le temps machine nécessaires dépendent à la fois du volume de données traitées et des choix propres à l'utilisateur en matière d'hyper-paramètres : taille des lots *(batch size)*, taux d'apprentissage *(learning rate)*, nombre d'époques *(epochs)*, etc.

5. Installé au premier semestre 2019 et inauguré le 24 janvier 2020, le supercaculateur Jean Zay du CNRS est actuellement le plus puissant de France, permettant 28 milliards d'opérations à la seconde, soit 28 PFlop/s (https://www.cnrs.fr/fr/jean-zay-le-supercalculateur-le-plus-puissant-de-france-pour-la-recherche). C'est grâce à ce supercalculateur qu'a été entraîné le modèle français FlauBERT (voir p. 66).

1.3.2 Caractérisation de la langue générale et de la langue de spécialité

Alors que les approches statistiques sont largement utilisées en TAL, j'ai exploré le bénéfice d'utiliser des indices linguistiques pour capturer des informations précises en domaine de spécialité, notamment l'identification de liens de causalité pour identifier des effets secondaires [Morlane-Hondère et al., 2015a], ou les verbes introduisant des noms de médicaments [Morlane-Hondère et al., 2015b] ou suggérant un mésusage médicamenteux [Campillos-Llanos et al., 2019]. J'ai commencé à aborder la problématique de l'impact du français inclusif sur les outils du TAL, en me focalisant sur l'étiquetage en parties du discours et la lemmatisation [Grouin, 2022]. Tous ces éléments linguistiques permettent de caractériser la langue générale et les langues de spécialité, en vue d'être réinjectés dans les approches statistiques.

Si les voisins distributionnels permettent de détecter des variantes, notamment orthographiques [Morlane-Hondère et al., 2016a], nous avons proposé une approche de normalisation de concepts fondée sur les différents parcours possibles de voisins distributionnels générés par word2vec, permettant soit d'identifier rapidement des candidats pertinents, soit davantage de candidats mais également davantage de candidats non pertinents [Morlane-Hondère and Grouin, 2016].

1.3.3 Utilisation raisonnée des modèles transformers

Si les modèles pré-entraînés (CamemBERT, FlauBERT) obtiennent des résultats élevés sur plusieurs tâches et bien qu'ils encodent des stéréotypes de genre comme nous l'avons observé sur des pièces du théâtre classique [Benamar et al., 2022b], j'ai exploré comment enrichir ces modèles en limitant le coût d'un réentraînement ou d'un affinage. Nous avons démontré que l'ajout d'une sous-couche d'informations de parties du discours suffit pour améliorer les performances sur trois tâches (reconnaissance d'entités nommées, implications textuelles, et paraphrases) en anglais et en français [Benamar et al., 2021], tandis que l'ajout d'une sous-couche d'entités nommées est nécessaire pour l'analyse de sentiments en anglais. L'ajout d'informations de locutions adverbiales, de valence, ou de sémantique devrait poursuivre l'amélioration des résultats. De plus, nous avons observé que les tokenizers utilisés par ces modèles (BPE, WordPiece) ont un impact sur le traitement des mots hors vocabulaires, avec une incidence plus forte sur les termes de spécialité que sur les fautes de frappe, et que les informations morpho-syntaxiques, parce qu'elles sont contextuelles, sont utiles pour le traitement des mots hors vocabulaire [Benamar et al., 2022c].

1.3.4 Application du traitement automatique des langues pour l'épidémiologie

Une collaboration avec Santé Publique France a porté sur la classification automatique des causes de décès parmi 40 regroupements syndromiques [6] [Baghdadi et al., 2019c], dans un objectif épidémiologique d'alerte sanitaire en temps réel (travaux non présentés dans ce manuscrit). Les résultats obtenus dépassent 0,94 de F-mesure quelle que soit l'approche testée (règles ou SVM) [Baghdadi et al., 2019a]. Pour comparer les performances des deux approches testées et dans un objectif de surveillance syndromique, chaque regroupement a été réparti parmi trois ensembles, selon que les valeurs de rappel, précision et F-mesure de ce regroupement se situent toutes trois dans le même intervalle (parmi quatre : ≥0,95, [0,90–0,95[, [0,85-0,90[, et <0,85) sur les deux approches, pour une seule approche, ou que les résultats sont hétérogènes [Baghdadi et al., 2019b]. La corrélation des informations extraites des certificats de décès avec les informations d'admissions aux urgences sur la même période nous permet de constater que les méthodes employées conservent les informations de variation saisonnière, tels que les épisodes de grippes en hiver, sans que les systèmes ne soient perturbés par ces variations.

6. Un regroupement syndromique *(Mortality Syndromic Group)* est un ensemble de causes médicales de décès correspondant à un code CIM-10 (pathologies, syndromes, symptomes). Ils couvrent différents domaines cliniques : cancer, problèmes respiratoires, maladies infectieuses, problèmes digestifs, problèmes cardiaques, etc.

1.3.5 Organisation de campagnes d'évaluation

Depuis 2007, une partie de mes activités porte sur l'organisation de campagnes d'évaluation, notamment le Défi Fouille de Texte[7] (DEFT) qui est proposé chaque année en français, sur des tâches et des domaines d'application régulièrement renouvelés, et dont l'atelier de clôture se tient pendant la conférence francophone TALN. Les corpus utilisés ont concerné des copies d'étudiants (2021 et 2022), des cas cliniques (2019 à 2021), des tweets (2015, 2017 et 2018), des nouvelles littéraires courtes (2014), des recettes de cuisine (2013), des articles scientifiques en sciences humaines (2011 et 2012), des articles de journaux (2008 à 2011), des transcriptions de débats parlementaires (2007 et 2009) et des pages Wikipédia (2008). Les tâches comprennent la correction automatique (copies d'étudiants), le classement (cas cliniques, nouvelles littéraires, recettes, pages Wikipédia), l'extraction d'information (cas cliniques, tweets, recettes), la fouille d'opinion (tweets, débats parlementaires), et la variation diachronique et diatopique (presse). J'ai également participé à l'organisation de tâches de reconnaissance de concepts cliniques et de codage CIM-10 de ces concepts en français pour le challenge international CLEF e-Health.

1.4 Contexte et organisation des travaux

Parce que chaque locuteur d'une langue ne produit jamais deux fois la même phrase dans le cadre d'une réalisation en discours [Gross, 1981], conclure qu'un problème donné est résolu (mon premier constat), et vouloir capturer l'intégralité de la sémantique des mots d'une langue dans un modèle distributionnel semblent improbables. Dans ce manuscrit, je me suis donc intéressé au rôle des utilisateurs dans les productions langagières, dans la mesure où ces productions peuvent servir, soit de matériau de base pour constituer de nouvelles ressources pour le TAL (tels que les modèles de langue), soit d'objet d'étude pour lequel des méthodes du TAL seront nécessaires (tels que les messages sur les réseaux sociaux).

Faisant suite à cette introduction générale (p. 7), mon manuscrit est organisé en deux parties. Dans la première partie, je reviens sur l'impact des locuteurs sur l'objet d'étude, tandis que la deuxième partie traite de l'impact des utilisateurs sur les outils et ressources du traitement automatique des langues. Ce manuscrit se termine par un chapitre de conclusion générale et de perspectives de recherche (p. 81).

1.4.1 Première partie, impact des locuteurs sur l'objet d'étude

Dans cette partie, le chapitre 2 est consacré à la production des utilisateurs. Je m'intéresse principalement aux productions langagières réalisées par les locuteurs sur les réseaux sociaux, et dans une moindre mesure, les productions orales, sur la base de transcriptions automatiques de la parole. J'étudie ces productions en mettant en avant une hétérogénéité linguistique d'abord (vocabulaire, style, complexité sémantique), et une hétérogénéité d'usage (émojis, hashtags, français inclusif) d'autre part.

Le chapitre 3 porte sur les non-dits des utilisateurs, que j'ai principalement étudiés à l'occasion de la thèse de Liyun Yan, qui a d'abord proposé une classification des inférences en s'inspirant des travaux de Peirce, qu'elle a ensuite appliquée lors d'une procédure d'annotation de corpus, avant d'utiliser ces informations pour effectuer une fouille d'opinion sur un corpus d'avis de touristes chinois en visite à Paris. Outre les spécificités du mandarin standard pour le TAL (plusieurs systèmes d'écriture, absence d'espace), nous concluons que l'identification et le typage des inférences constituent une piste complémentaire à la fouille d'opinion plus traditionnelle, pour tenir compte des usages langagiers propres à chaque locuteur.

7. DEFT (DÉfi Fouille de Texte) : https://deft.lisn.upsaclay.fr/

1.4.2 Deuxième partie, impact des utilisateurs sur les outils et ressources du traitement automatique des langues

La deuxième partie comprend également deux chapitres. Je consacre le chapitre 4 au domaine médical et plus précisément les approches que j'ai explorées pour les activités de pharmacovigilance sur les réseaux sociaux, dans le cadre du stage de Dalia Megahed et des post-doctorats de François Morlane-Hondère puis Leonardo Campillos-Llanos. Dans un premier temps, je reviens sur les essais de modélisation des informations d'effets secondaires médicamenteux que j'ai étudiés avec Dalia et François sur des témoignages rapportés sur les réseaux sociaux, puis sur la recherche d'indices linguistiques pour améliorer la détection d'effets secondaires (verbes introducteurs de médicaments, analyse de voisins distributionnels, etc.), avant de conclure sur une preuve de concept autour de la complexe détection du mésusage, réalisée avec Leonardo.

Dans le chapitre 5, je traite des expériences menées à base de modèles distributionnels, notamment l'impact de la segmentation en sous-unités sur les mots hors vocabulaire et l'étude des stéréotypes de genre dans les pièces de théâtre classique que j'ai explorés pendant la thèse d'Alexandra Benamar, puis la recherche d'une solution alternative au réentraînement et à l'affinage vers laquelle j'ai orienté Alexandra qui a combiné des informations de parties du discours dans les représentations vectorielles. Malgré les difficultés pour traiter la parole transcrite, je présente les travaux effectués en stage par Lufei Liu sur la segmentation thématique automatique de transcription de journaux d'information télévisés.

Première partie

Impact des utilisateurs sur l'objet d'étude

Chapitre 2

La production des utilisateurs

Sommaire

2.1 Introduction

Depuis le « web 2.0 » qui met l'utilisateur au centre des actions et le développement d'applications dites « participatives » [O'Reilly, 2005], les forums de discussion puis les micro-blogs à base de messages textuels (en particulier Twitter [1] et Reddit [2] lancés entre 2005 et 2006) constituent une source d'information utile dans de nombreux domaines mais critiquée pour l'absence de contrôle [3]. Dans le domaine médical, des forums dédiés à la santé [4] permettent aux patients putatifs ou avérés de chercher des informations sur leur maladie, de se constituer en groupe de malades, et d'échanger entre eux. Il a été démontré qu'un patient qui connaît sa maladie, qui s'informe correctement (« littératie de santé » [5]) [Dib et al., 2022] et qui est pleinement acteur de son processus de soin est plus apte à suivre son traitement, notamment dans le cadre des maladies chroniques [Lorig et al., 2001, Ye et al., 2022], ce qui conduit, en retour, les professionnels de la santé à développer une démarche centrée sur le patient [6]. En dépit d'une dimension négative, les réseaux sociaux offrent un corpus en permanence renouvelé [7] pour la recherche en traitement automatique des langues : analyse de la crédibilité des utilisateurs [Hassan et al., 2018], analyse de la réputation et de l'impact de la communication d'une entreprise [Araujo and Kollat, 2018], prédiction des résultats d'élections [Wang and Gan, 2017], etc. Les contenus produits par les utilisateurs sur les réseaux sociaux répondent à un impératif d'immédiateté et de brièveté généralement imposé par le réseau social, en vertu d'une monétisation du volume de consultations et d'actions

1. https://twitter.com/home
2. https://www.reddit.com/fr/
3. Chaque utilisateur peut aisément témoigner sur le sujet de son choix par un message, des photos, une vidéo, mais également inventer des témoignages ou contribuer à relayer des rumeurs sans prise de recul ou de manière totalement assumée pour les utilisateurs qualifiés de « trolls ».
4. Le leader français Doctissimo (https://www.doctissimo.fr/) propose ainsi un forum général dédié à la santé (https://forum.doctissimo.fr/) et des sous-forums dédiés à certaines conditions tel celui sur la grossesse et les nourrissons (https://forum.doctissimo.fr/grossesse-bebe/liste_categorie.htm) ou catégories d'information, tels que les médicaments (https://forum.doctissimo.fr/medicaments/liste_categorie.htm).
5. https://www.santepubliquefrance.fr/docs/la-litteratie-en-sante-un-concept-critique-pour-la-sante-publique
6. https://www.has-sante.fr/upload/docs/application/pdf/2015-06/demarche_centree_patient_web.pdf
7. On estime actuellement que 6000 tweets sont envoyés chaque seconde dans le monde, avec des pics en cas d'événements comme lorsque la France marque un but pendant la coupe du monde de football, occasionnant 24 400 tweets à la seconde selon Elon Musk. D'après l'institut de sondage Médiamétrie, Twitter est le 38ème site le plus visité en France, avec 5 329 000 visiteurs uniques chaque jour et 16 874 000 visiteurs uniques chaque mois en moyenne (données juin 2022, https://www.mediametrie.fr/fr/audience-internet-global-en-france-en-juin-2022), pour environ 10 millions d'utilisateurs actifs en janvier 2022 (https://www.blogdumoderateur.com/chiffres-twitter/).

sur ces contenus (action d'aimer un message et de rediffuser le message à ses propres abonnés [8]). Toutes ces productions rapides pour une consultation dans l'immédiateté sont bruitées. Depuis 2015, le workshop scientifique W-NUT [9] (Workshop on Noisy User-generated Text) se focalise sur les méthodes de TAL appliquées aux contenus bruités, notamment issus des réseaux sociaux. J'ai pu aborder la problématique des productions d'utilisateurs sur les réseaux sociaux dans le cadre de la thèse de Liyun Yan sur la fouille d'opinion en chinois, et durant les post-doctorats de François Morlane-Hondère et de Leonardo Campillos-Llanos sur la pharmacovigilance, puis pendant la pandémie de Covid-19, pour détecter l'auto-médication rapportée sur Twitter. J'ai également régulièrement travaillé sur des corpus de tweets pour les campagnes d'évaluation francophones DEFT [10] (DÉfi Fouille de Texte) que je co-organise depuis 2007. Cela concerne en particulier les éditions 2015 sur la fouille d'opinion, sentiment, et émotion sur la thématique du changement climatique [Hamon et al., 2015], 2017 sur l'analyse d'opinion et du langage figuratif [Benamara et al., 2017], et 2018 autour de la recherche d'information et l'analyse de sentiments sur la thématique des transports en Ile-de-France [Paroubek et al., 2018].

Par opposition, il existe également des ressources créées par une communauté d'utilisateurs spécialisés dans un domaine, tel l'astrophysique, qui reposent sur la contribution de ces spécialistes. Dans le domaine de l'observation de phénomènes transitoires (caractérisés par des émissions de courte durée comme les explosions de supernova, les sursauts radio rapides, et les sursauts gamma), plusieurs plateformes existent pour compiler les différents rapports d'observation, telles que AstroNote [11], Circulaires [12] et Astronomer's Telegram [13] [Rutledge, 1998]. Des applications ont été développées pour traiter ces contenus. C'est notamment le cas de la plateforme Astro-Colibri [14] [Reichherzer et al., 2021], développée à l'institut de recherche sur les lois fondamentales [15] au CEA à Saclay, pour analyser en temps réel les messages générés par les différents observatoires du monde, spécialisée dans l'analyse des phénomènes transitoires. J'explore le besoin d'outils en TAL pour l'analyse des rapports d'observation pour ces phénomènes d'astrophysique pendant la thèse d'Atilla Alkan.

Dans ce chapitre, je m'intéresse à présenter les différents phénomènes qui se retrouvent dans les productions numériques de locuteurs humains sur les réseaux sociaux, dans l'objectif de traiter ces phénomènes et de mesurer leur impact sur les approches de TAL. Pour cela, je fais une distinction entre l'hétérogénéité au niveau linguistique (section 2.2) et l'hétérogénéité directement liée aux usages des utilisateurs (section 2.3). Dans le premier cas, l'hétérogénéité linguistique peut s'imposer aux locuteurs (problématique de translittération des mots étrangers et de choix de vocabulaire, p. 18). Elle se retrouve en corpus en fonction de l'origine géographique du locuteur (l'opposition entre les sinogrammes traditionnels et simplifiés apparaît sur les réseaux sociaux en chinois, p. 18). Mais l'hétérogénéité peut également être un choix conscient des locuteurs, comme pour le *code switching* (p. 19) ou l'usage de certaines figures de style telles que le chleuasme (p. 19). J'aborde également la difficulté de l'interprétation sémantique, au niveau de l'assertion des informations fournies par un utilisateur (p. 22). L'évolution des usages sur les réseaux sociaux a conduit les utilisateurs à s'emparer des nouvelles fonctionnalités offertes telles que l'intégration d'émojis (p. 24) ou de hashtags (p. 25), et à en jouer dans leurs productions numériques. Pour le TAListe, la prise en compte de ces fonctionnalités constitue les deux faces d'une même pièce :

8. Cette action est nommée « retweet » sur Twitter, « reblog » sur d'autres micro-blogs.
9. http://noisy-text.github.io/2022/
10. DEFT : https://deft.lisn.upsaclay.fr/
11. AstroNote : https://www.wis-tns.org/astronotes/
12. Circulaires : https://gcn.nasa.gov/circulars et https://gcn.gsfc.nasa.gov/gcn3_archive.html pour les archives.
13. Astronomer's Telegram (ATel) : https://www.astronomerstelegram.org/
14. Astro-Colibri : https://irfu.cea.fr/Phocea/Vie_des_labos/Ast/ast.php ?t=fait_marquant&id_ast=4959 et https://astro-colibri.com/ pour l'accès à l'application.
15. Institut de recherche sur les lois fondamentales (IRFU) : https://irfu.cea.fr/

ces fonctionnalités sont utiles pour les tâches classiques de TAL (classification, repérage d'entités nommées, fouille d'opinion), mais elles appellent des traitements dédiés. Enfin, les locuteurs les plus militants ou les plus politisés peuvent reprendre certains codes, comme dans le cas du français inclusif (p. 27), quitte à combiner les codes linguistiques du français inclusif avec les usages spécifiques des réseaux sociaux en créant des hashtags en français inclusif (p. 28).

J'écarte de ce chapitre les productions d'utilisateurs que nous avons constituées à l'occasion de certaines éditions de DEFT, mais qui correspondent à des normes fixes, tant en termes d'organisation du contenu (recettes de cuisine [Grouin et al., 2013b]) que de style littéraire (articles scientifiques [Grouin et al., 2011b, Paroubek et al., 2012], et nouvelles courtes [Hamon et al., 2014b]).

2.2 Hétérogénéité linguistique

La langue est un objet vivant, qui appartient aux locuteurs de cette langue en ce sens qu'ils la modèlent et la font vivre par son utilisation. Bien que l'on désigne une langue donnée par un singulier, « le français », il existe plusieurs langues en fonction des usages. On oppose généralement les langues de spécialité, propres à un domaine particulier (juridique, médical, scientifique, etc.), à la langue dite « générale » utilisée par tous les locuteurs dans leur usage quotidien. Même pour une langue de spécialité, il existe une variété de genres ou de sous-langues de spécialité, avec des caractéristiques spécifiques. Dans le domaine médical, on observe une différence entre la langue utilisée dans les dossiers patients et celle employée dans les articles scientifiques [Friedman et al., 2002]. Cette variété offre l'opportunité d'inventorier les différents genres de texte lors de la création de corpus spécialisés [Zweigenbaum et al., 2001]. Plus spécifiquement, les langues de spécialité se caractérisent par un vocabulaire qui leur est propre, mais également par une distribution différente des marqueurs linguistiques, telle que la négation [Cohen et al., 2017].

2.2.1 Hétérogénéité de vocabulaire

À l'occasion des travaux de thèse de Liyun sur la fouille d'opinion sur les réseaux sociaux en chinois [Yan, 2021], nous avons pu mettre en évidence une hétérogénéité linguistique au niveau des mots. Cette hétérogénéité s'explique par au moins deux aspects.

L'hétérogénéité linguistique s'explique d'abord par la langue utilisée par les utilisateurs des réseaux sociaux. Parce que l'objet d'étude concerne les avis de touristes chinois en visite à Paris, la langue employée dans les avis est le mandarin standard, langue officielle en Chine. Si cette langue est la même entre la Chine, Singapour et Taïwan, le système d'écriture utilisé diffère : les sinogrammes simplifiés sont en usage en Chine et à Singapour, alors que les sinogrammes traditionnels ont été conservés à Hong Kong, Macao et Taïwan. Parce que l'usage des réseaux sociaux n'est pas limité aux seuls chinois de Chine, le corpus étudié combine des messages rédigés dans les deux systèmes d'écriture, ce qui constitue une problématique forte pour le TAL du chinois : soit les outils devront gérer les deux systèmes d'écriture, soit une conversion d'un système d'écriture vers un autre (généralement du traditionnel vers le simplifié, ou vers un système alternatif comme le pinyin) devra être envisagée.

L'hétérogénéité linguistique s'explique également par la thématique abordée dans le corpus, en l'occurrence le tourisme à Paris, ce qui implique des références culturelles françaises différentes de celles existantes en Chine, avec une problématique autour des entités nommées et de la manière de les écrire. Plus particulièrement pour la thématique du tourisme, cela concerne les noms de lieux (villes, monuments, transports, magasins, etc.) et de personnes. La translittération des mots français dans le corpus chinois peut se faire de trois manières : soit par le biais de sinogrammes, soit au moyen du système d'écriture pinyin (fondé sur l'alphabet latin), soit en intégrant directement le mot français dans le texte. Le choix d'une solution repose à la fois sur la possibilité linguistique d'une translittération (a priori impossible pour les acronymes), sur la popularité de l'entité à translittérer (un consensus existe uniquement pour les entités connues comme les monuments touristiques populaires), et sur la confiance du locuteur dans sa capacité à faire le bon choix.

Cette double problématique au niveau linguistique (coexistence de plusieurs systèmes d'écriture et translittération des mots étrangers) s'ajoute à celles déjà connues au niveau informatique comme la tokénisation des séquences de sinogrammes qui s'écrivent sans espace, ou la représentation du contenu par des plongements de mots.

Translittération des mots étrangers

L'une des particularités observée dans ce corpus d'avis de touristes chinois à Paris concerne la translittération des mots français, en particulier pour les lieux visités par les touristes et pour les éléments de la vie quotidienne (transports en commun, magasins, etc.). Si une correspondance phonétique ou sémantique existe, elle sera utilisée mais pourra donner lieu à des variantes (les translittérations de l'exemple 1 renvoient toutes aux Champs-Élysées, parmi onze variantes identifiées en corpus). Faute de correspondance, ou s'il n'existe pas de suite de sinogrammes attestée par les locuteurs, le mot français pourra être intégré directement dans le message (exemple 2).

(1) 香榭里舍大道 (xiāng xiè lī shè dàdào). *Champs-Élysées*
香榭里舍大街 (xiāng xiè lī shè dàjiē). *Champs-Élysées*
香榭利舍大街 (xiāng xiè lì shè dàjiē). *Champs-Élysées*

(2) 位置 超级 方便 ，就 在 vendome 广场 边上 (wèizhì chāojí fāngbiàn, jiù zài vendome guǎngchǎng biān shàng) *L'emplacement est super pratique, juste à côté de la place Vendôme*

Les emprunts concernent trois cas : les marques peu ou pas connues en Chine et pour lesquelles n'existent pas de version en sinogrammes (exemple 3), les sigles (exemples 4), et les noms propres (exemple 5), en raison de l'impossibilité de trouver une correspondance phonétique d'une part, et de la présence de ces éléments dans la signalétique sur le terrain pour les marques et sigles d'autre part. Nous constatons que trois marques sont mentionnées dans l'exemple 3 dont deux pour lesquelles existent des séquences de sinogrammes (Sephora : 丝 芙 兰, McDonald's :麦当劳) alors que la troisième est indiquée de manière condensée en alphabet latin (Marks & Spencer : marksSpencer), probablement en raison de sa faible implantation en Chine.

(3) 离 地铁站 很 近 附近 有 丝 芙 兰 有 麦当劳 marksSpencer 挺 方便 的 。(lí dìtiě zhàn hěn jìn fùjìn yǒu sīfùlàn yǒu màidāngláo marksSpencer tǐng fāngbiàn de) *C'est très pratique d'avoir Sephora, McDonald's et Marks & Spencer près de la station de métro.*

(4) 巴黎 里昂 火车站 有 地铁 1 号 和 14 号线 ，还有 多条 通往 巴黎 近郊 的 RER 轻轨 线路 ，可谓 四通八达 。(bālí lǐáng huǒchē zhàn yǒu dìtiě 1 hào hé 14 hào xiàn, hái yǒu duō tiáo tōng wǎng bālí jìnjiāo de RER qīngguǐ xiànlú, kěwèi sìtōngbādá). *La gare de Paris Lyon propose les lignes de métro 1 et 14, ainsi que plusieurs lignes de métro léger RER vers la banlieue parisienne, qui vont dans toutes les directions.*

(5) 员工 非常 热情 友善 ，尤其 是 Martin (yuángōng fēicháng rèqíng yǒushàn, yóuqí shì Martin) *Le personnel est très chaleureux et sympathique, surtout Martin*

Homogénéisation des systèmes d'écriture et normalisation des entités nommées

Pour contourner les problèmes liés à l'emploi de plusieurs systèmes d'écritures (sinogrammes simplifiés et sinogrammes traditionnels) dans le même corpus, un premier pré-traitement a consisté à convertir automatiquement le chinois traditionnel en chinois simplifié, au moyen de l'outil OpenCC [16]. D'autre part, en raison de la diversité des translittérations possibles d'entités nommées faisant référence à des mots non chinois (utilisation de sinogrammes, emprunts, etc.), et de la

16. Open Chinese Convert (OpenCC) : https://pypi.org/project/OpenCC/

difficulté à tokéniser correctement cette diversité avec l'outil Jieba [17], Liyun a converti toutes les variantes d'une même entité en pinyin [18], sans conserver les marqueurs de ton, comme préconisé par [Jiang et al., 2007]. Ainsi, toutes les variantes de l'entité « Champs-Élysées » (cf. exemple 1) sont rassemblées sous l'unique forme « *xiang xie li she* ». Il en résulte que ces formes uniques peuvent être ajoutées au dictionnaire interne de Jieba, ce qui permet d'améliorer la tokénisation et l'étiquetage en parties du discours.

2.2.2 Hétérogénéité de styles

Parce que la langue appartient à ses locuteurs, ces derniers ont également l'opportunité de personnaliser leurs productions par des formulations spécifiques. Ces formulations peuvent être liées à un besoin, comme dans le cas du *code switching* lorsqu'un concept est absent d'une langue, ou pour démarquer son discours de celui des autres locuteurs, en faisant usage de figures de style.

Code switching

Dans le corpus d'avis touristiques en chinois, nous observons plusieurs cas de *code switching* provenant essentiellement de l'anglais (mais également du coréen et du japonais). Dans la plupart des cas, il s'agit de changement de langue volontaire dans une recherche de convivialité (exemple 6). Plus rarement, cela concerne un terme technique absent de la langue ou une technologie qui n'a pas été inventée par les chinois, impliquant un changement de langue obligatoire, comme pour le Wi-Fi (exemple 7), technologie dont le nom correspond à une marque déposée, et qui n'a pour équivalent dans la langue qu'une expression plus générique « *réseau sans fil* » (无线网络) qui renvoie aussi bien aux réseaux informatiques que télécoms, même si les jeunes générations interprètent d'emblée cette séquence comme renvoyant au wifi.

(6) 铁塔 view 很 不错 (tiětǎ view hēn bùcuò) *La « view » sur la Tour est très bien*

(7) 房间内 wifi 几乎 无法 使用 (fángjiān nèi wifi jīhū wúfǎ shǐyòng) *Le « wifi » est presque inutilisable*

Nous observons cependant qu'il est plus rare que ce phénomène porte sur une phrase complète, sauf pour conclure son avis de manière internationale (exemple 15). Ce *code switching* est volontaire puisque le locuteur n'est pas contraint par une absence de concept dans sa langue, mais que cela correspond à une recherche de style personnel pour appuyer son propos.

(8) 安保 工作 做得好 ，可以 借 吹风机 。good location ！ (ānbǎo gōngzuò zuò dé hǎo, kěyǐ jiè chuīfēngjī. Good location !) *Sécurité bien faite, possibilité d'emprunter un sèche-cheveux. « Good location ! »*

Littérature de Versailles

En novembre 2020, une millionaire et influenceuse chinoise a raconté un épisode de sa vie sur le réseau social chinois Weibo, sans se rendre compte du décalage entre sa vie et celle des autres chinois [19]. Les internautes ont critiqué cette vie luxueuse et superficielle qu'ils ont associée à celle

17. L'outil Jieba effectue une tokénisation du chinois en s'appuyant sur l'algorithme de Viterbi et sur un dictionnaire interne de préfixes, qu'il est possible de compléter, aussi bien avec des préfixes qu'avec des entités translittérées complètes. Cependant, toute forme absente du dictionnaire engendrera de possibles erreurs de tokénisation, en particulier pour les translittérations. L'impossible exhaustivité des dictionnaires se pose pour les entités nommées de manière générale, et plus encore pour la translittération d'entités nommées étrangères. Cet outil permet également un étiquetage en parties du discours. https://github.com/fxsjy/jieba

18. Il existe des modules développés en Python qui permettent cette conversion automatique, tel que le module pinyin : https://pypi.org/project/pinyin/

19. « Il n'y a pas assez de bornes de recharge pour voitures électriques dans le quartier, et nous ne sommes pas autorisés à en installer d'autres. Nous n'avons pas eu d'autre choix que de déménager dans une plus grosse maison avec garage privé pour la Tesla de mon mari. » (traduction personnelle à partir d'une traduction en anglais disponible sur https://pandaily.com/versailles-literature-trending-on-chinas-internet-a-new-way-to-brag/).

de la cour de Versailles. Pour tourner en dérision ce témoignage et dénoncer les millionnaires déconnectés de la réalité, un concours de « littérature de Versailles » (凡尔赛文学 – fán'ěrsài wénxué) a été lancé. L'expression, tirée du manga *The Rose of Versailles*, a ainsi été utilisée pour qualifier, sur les réseaux sociaux, les messages de fausse modestie dans lesquels le locuteur se rabaisse ou se plaint de manière explicite dans l'objectif de se vanter implicitement [Ren and Guo, 2021].

Dans l'exemple 9, il importe de connaître les monuments parisiens et leur popularité touristique pour comprendre que l'émetteur du message ne déplore pas de ne pas voir l'Arc de Triomphe (monument de second rang), mais qu'il se vante de voir la Tour Eiffel scintiller depuis ses fenêtres. Le locuteur met en avant un faux mécontentement pour mieux affirmer son statut économique.

(9) 升级了最贵的顶楼套房，也无法看到凯旋门夜景，只能看到埃菲尔铁塔。(shēngjíle zui guì de dǐnglóu tàofáng, yē wúfǎ kàn dào kāixuánmén yèjǐng zhī néng kàn dào āifēi'ěr tiětǎ). *Alors que j'ai pris la suite penthouse la plus chère, je ne vois même pas l'Arc de Triomphe de nuit, seulement la Tour Eiffel.*

Ce type de figure de style rappelle le *chleuasme*, qui consiste à se déprécier pour attirer la sympathie du public, à ceci près que dans le cas de la littérature de Versailles, l'objectif n'est pas de s'assurer la sympathie du récepteur du message, mais uniquement de se vanter en feignant une critique. Plusieurs travaux ont étudié le chleuasme en littérature classique, notamment dans les œuvres de La Rochefoucauld, où cette « *griserie de l'éloge* » est une quête de compliments, par amour propre, au moyen de calculs conscients [Tourrette, 2012], ou encore en littérature contemporaine comme chez Françoise Sagan [Hromadova, 2019]. Les travaux sur les figures de style employées sur les réseaux sociaux semblent cependant plus rares. Au-delà de ces constats, aucun traitement n'a été réalisé autour de cette hétérogénéité de style, en raison de sa rareté dans le corpus d'avis touristiques, mais également parce qu'il s'agit d'une réalité de terrain que les méthodes de TAL ont à prendre en compte.

Les influenceurs : se distinguer pour monétiser son contenu

Un autre aspect de la révolution numérique concerne les « influenceurs », arrivés en 2010 par le biais des « youtubeurs ». Il s'agit d'utilisateurs actifs, parvenant à fédérer une communauté de « suiveurs » qui partagent leurs goûts ou leurs points de vue, puis qui profitent de cette notoriété pour la monétiser auprès de marques en communiquant sur les produits de ces marques. À l'occasion des vingt ans de la télé-réalité en France, nous avons analysé et comparé deux émissions (la première « télé-réalité » en France avec *Loft Story* diffusée en 2001 et l'émission *Les Marseillais à Dubaï* diffusée en 2021) et mis en évidence une évolution des expressions d'opinion, sentiment, émotion dans les productions langagières qui s'explique à la fois par l'évolution de la télé-réalité et par l'apparition de cette fonction d'influenceur [Biscarrat et al., 2022]. Alors que les participants de *Loft Story* ne se connaissaient pas, étaient célibataires et venaient pour trouver l'amour, les participants des *Marseillais à Dubaï* sont issus du monde de la nuit, se sont déjà croisés dans ces établissements, bénéficient de contrats d'acteur, et peuvent également être influenceurs en dehors de l'émission. Un point commun à ces émissions concerne le « confessionnal », dispositif où chaque participant est seul face à la caméra, et que nous avons privilégié pour deux raisons : l'absence de parole superposée (complexe pour les systèmes de transcription de la parole) et le rééquilibrage du temps de parole femmes/hommes (la parole est majoritairement masculine en dehors du confessionnal malgré un temps de présence équilibré femmes/hommes à l'écran). Nous avons appliqué le lexique Emotaix [Piolat and Bannour, 2009], constitué de 4921 termes organisés hiérarchiquement en émotions, sur la transcription du confessionnal (11 épisodes de *Loft Story* soit 26h, et 61 épisodes des *Marseillais à Dubaï* soit 60h). Le tableau 2.1 présente les principales émotions identifiées dans ces passages de confessionnal, classées par distribution décroissante dans la parole féminine de l'émission *Loft Story*.

Emotion	Classe d'émotion	Loft Story			Les Marseillais à Dubaï		
		Femme	Homme	Ratio F	Femme	Homme	Ratio F
Timidité	anxiété	4	0	1,00	1	1	0,50
Tristesse	mal-être	13	1	0,93	8	6	0,57
Douleur	mal-être	5	1	0,83	3	9	0,25
Apaisement	bien-être	4	1	0,80	12	5	0,71
Attirance	bienveillance	4	1	0,80	6	6	0,50
Désir	bienveillance	11	4	0,73	15	17	0,47
Amour	bienveillance	26	10	0,72	28	21	0,57
Pleur	mal-être	2	2	0,50	0	1	0,00
Joie	bien-être	1	1	0,50	3	4	0,43
Plaisir	bien-être	2	6	0,25	5	19	0,21
Insatisfaction	mal-être	7	3	0,70	6	0	1,00
Humiliation	mal-être	0	1	0,00	1	0	1,00

TABLE 2.1 : Nombre d'émotions principales identifiées dans la parole féminine et masculine sur des passages de confessionnal, par distribution décroissante dans la parole féminine du Loft

Nous constatons que les classes d'émotion identifiées majoritairement dans la parole féminine sur les deux émissions correspondent essentiellement à l'anxiété ou au mal-être (« *timidité* », « *tristesse, douleur* » dans *Loft Story*, correspondant au domaine de la romance, « *insatisfaction, humiliation* » pour *Les Marseillais à Dubaï*, relevant davantage d'une compétition), les émotions de bien-être étant davantage partagées entre femmes et hommes. Si les émotions de bienveillance (« *amour, attirance, désir* ») étaient davantage véhiculées par les femmes dans *Loft Story*, elles sont rééquilibrées entre femmes et hommes dans *Les Marseillais à Dubaï*. Alors qu'on pourrait y voir une évolution favorable, il semble que cela correspond en réalité à une appropriation par les hommes du « *money shot* » (séquence qui fait de l'audience avec une valeur marchande) en raison de la monétisation possible dans le futur en tant qu'influenceur.

L'application de méthodes du TAL écrit sur de la parole transcrite reste cependant complexe. La télé-réalité produit de la parole spontanée, caractérisée par des hésitations ou répétitions, et l'emploi de termes propres à chacun. Au niveau technique, les sorties de systèmes de transcription de la parole sont généralement en minuscules et sans signe de ponctuation. Par ailleurs, la qualité de la transcription fluctue fortement, occasionnant des erreurs pouvant conduire à des contre-sens.

Le tableau 2.2 présente les transcriptions automatiques des personnages de Paga (homme)

Personnage	Transcription	
Paga	Automatique	françois lenglet de ce soir avec l' ena peut faire avancer un petit peu ma marre concernant le sens euh quelle décision je prends il faut toujours être nouveau style donc jamais
	Manuelle	*Je pense que le « date » de ce soir avec Léna, peut faire avancer un petit peu ma, ma pensée à moi, dans le sens heu, quelle décision je prends. Il faut toujours être beau, sait-on jamais ?*
Victoria	Automatique	il a il pas honnête si on veut y joue venner et j' en peux plus et moi enfin j' ai envie qu' il a su tout je suis ravi qu' il assume que il y a encore entre autres
	Manuelle	*Ilan, il est pas honnête, et en fait il joue avec mes nerfs, et j'en peux plus. Et moi en fait j'ai envie qu'il assume tout. J'ai envie qu'il assume qu'il y a encore un truc entre nous.*

TABLE 2.2 : Transcriptions automatiques et manuelles d'extraits des Marseillais à Dubaï

et Victoria (femme) par le système d'ASR développé au LIUM sur l'émission *Les Marseillais à Dubaï*, ainsi que la transcription manuelle que j'ai faite en écoutant l'émission. Nous observons des erreurs dans la transcription automatique qui peuvent s'expliquer par les raisons suivantes :

— le modèle de reconnaissance de la parole a été entraîné sur du contenu journalistique, complété d'un dictionnaire d'entités nommées pour améliorer les performances sur les noms propres ; toute nouvelle entité absente du dictionnaire sera mal transcrite (la pandémie « Covid-19 » est transcrite « *koweït dix-neuf* » dans les émissions d'information)

— l'émission a été montée en intégrant un fond musical dans 93% des plans, contribuant à bruiter le signal et dégrader les performances de l'ASR

— la prosodie et les termes employés par les personnages de l'émission sont représentatifs de la parole spontanée, à l'opposé de la parole préparée utilisée dans les émissions d'information

La combinaison de ces propriétés explique l'entité nommée « François Lenglet » en remplacement de la séquence « *Je pense que le "date"* » prononcée par Paga (sur fond musical à un niveau élevé, doublé d'un *code switching*) et la séquence finale « *être nouveau style donc jamais* » au lieu de « *être beau, sait-on jamais ?* » (prosodie marquée et voix plus faible au début de l'interrogation). Compte tenu de ces observations, le résultat d'une extraction d'information et d'une fouille d'opinion sur des sorties ASR de ce niveau de qualité apparaît particulièrement complexe. Les analyses en sciences humaines fondées sur les descripteurs extraits automatiquement par des approches de TAL doivent tenir compte de ces difficultés d'ordre informatique.

2.2.3 Complexité sémantique

Il existe également une difficulté d'ordre sémantique relative à la diversité des informations renseignées par les utilisateurs. J'ai été confronté à cette diversité dans mes travaux de pharmacovigilance sur les réseaux sociaux (voir section 4.1) et lors de participation à la campagne d'évaluation i2b2 2009 et 2010 sur l'extraction d'informations médicamenteuses et les assertions autour de ces médicaments depuis des comptes-rendus cliniques en anglais [20]. Si les complexités d'interprétation sémantique ne sont pas spécifiques au domaine médical, elles ont cependant des conséquences potentiellement plus importantes en raison des enjeux de santé associés. Dans un objectif de surveillance des effets secondaires produits par les médicaments, la principale difficulté consiste à déterminer si le symptôme rapporté est causé par un seul médicament ou s'il résulte d'une interaction médicamenteuse. L'exemple 10 est un témoignage posté sur un forum de santé grand public, qui mentionne quatre traitements *(Tramadol, Naprosyne, Diclofénac, Oxycontin)* et rapporte des effets secondaires de type *douleurs* et *prurit*. Ces douleurs sont-elles causées par un seul des traitements mentionnés, par l'interaction de plusieurs d'entre eux, ou encore par un élément non-mentionné dans ce témoignage (un traitement régulier, une consommation excessive d'alcool ou de café) ?

(10) Maintenant de nouveau forte douleur au dos. D'abord Tramadol puis Naprosyne et Diclofénac et de nouveau Oxycontin. Cette nuit énormément de douleurs, impossible de rester couché, assise, debout, rien à faire. Impossible de dormir à cause de la démangeaison, vraiment un très désagréable effet secondaire, comme si des bestioles me courraient sur la peau.

Sur la campagne i2b2 2010, après avoir identifié les pathologies, il fallait également déterminer l'assertion par rapport à cette pathologie, parmi six possibilités (présent, absent, possible, hypothétique, conditionnel [21], associé à quelqu'un d'autre) [Uzuner et al., 2011]. Sur l'exemple 11, la pathologie *heart disease* (maladie cardiaque) ne concerne pas directement le patient mais fournit une information sur la cause du décès des parents, et constitue vraisemblablement pour les

20. https://www.i2b2.org/NLP/Medication/ (2009) et https://www.i2b2.org/NLP/Relations/ (2010).

21. Une pathologie est *conditionnelle* si elle ne se produit que sous certaines circonstances (problèmes hépatiques si la dose de 7g de paracétamol par jour est dépassée), *hypothétique* si elle peut se produire dans le futur (existence d'un cancer chez les parents du patient, facteur héréditaire), et *possible* si de premiers éléments le laissent entrevoir.

médecins un risque héréditaire pour ce patient. Au point de vue TAL, il fallait donc associer à cette pathologie l'assertion *associé à quelqu'un d'autre*. Les maladies de l'exemple 12 sont hypothétiques puisque le patient ne les subit pas actuellement alors que l'hyperglycémie (exemple 13) est possible d'après les résultats d'examen. Enfin, la dyspnée dans l'exemple 14 est conditionnelle puisqu'elle ne se produit pas en permanence mais qu'elle est liée à la montée des escaliers.

(11) Father and mother died in their eighties of heart disease. *(Le père et la mère sont décédés de maladie cardiaque au-delà de 80 ans)*

(12) Please return to the Deanna if you experience SOB, chest pain, fevers, chills, dizziness, decreased urine output. *(Reprendre du Deanna en cas d'essoufflements, douleurs thoraciques, fièvre, frissons, étourdissements, et diminution de la production d'urine)*

(13) She is at approximately 40% risk of having hyperglycemia. *(Elle est approximativement à 40% de risque d'être en hyperglycémie)*

(14) Three days ago, he developed dyspnea on exertion when climbing stairs, assoc w/ 2-pillow orthopnea and PND. *(Il y a trois jours, il a développé une dyspnée à l'effort en montant les escaliers, associée à une orthopnée mesurée à deux oreillers et une dyspnée nocturne paroxystique)*

Nous avons mis en place deux systèmes, l'un à base de SVM en considérant l'identification de l'assertion comme une tâche de classification, l'autre à base de règles dérivées de NegEx [Chapman et al., 2001]. C'est par l'approche à base de SVM que nous avons obtenus nos meilleurs résultats [Minard et al., 2011], y compris sur la classe « *associé à quelqu'un d'autre* » (avec des valeurs de rappel à 0,779, une précision de 0,856 et une F-mesure de 0,816). En revanche, l'identification des classes *possible* et *conditionnel* s'est révélée complexe au regard des résultats obtenus.

Dans le domaine médical, compte-tenu des enjeux de santé, il importe de pouvoir déterminer avec succès quelles sont les informations valables au moment présent. Il est donc essentiel de discriminer les différentes pathologies mentionnées, et pour la pharmacovigilance, de distinguer les traitements actuels parmi les différents traitements listés. Ces distinctions sémantiques s'appuient, en partie, sur une analyse temporelle du contexte, comme celle que nous avons mise en place lors de notre participation à la campagne SemEval en 2016 [Grouin and Moriceau, 2016].

2.2.4 Des choix individuels qui conditionnent les approches de TAL utilisées

Dans cette section, nous avons étudié l'impact des choix des locuteurs, en termes de vocabulaire (essentiellement sur les mots étrangers) et de système d'écriture (lié à l'origine géographique du locuteur) d'une part et de style (*code switching* volontaire, figure de style tel que le chleuasme) d'autre part. Les premières observations sont dépendantes de la langue (en l'occurrence le mandarin standard), et de la thématique abordée (le tourisme à Paris). La langue conditionne le système d'écriture tandis que la thématique traitée a pour conséquence de rendre complexe certaines traductions de mots étrangers, en particulier des concepts et des entités nommées, impliquant parfois une hétérogénéité de vocabulaire. Au-delà des spécificités propres au traitement automatique du chinois, nous avons étudié l'opportunité d'homogénéiser les systèmes d'écriture (vers des sinogrammes simplifiés ou vers du pinyin intégral) et de normaliser les entités nommées étrangères. Nous avons également vu qu'il existe une difficulté liée à l'interprétation sémantique du contenu des productions langagières pour distinguer, au niveau des assertions, si une information mentionnée est toujours d'actualité. Le traitement de la parole transcrite, quelle que soit la tâche effectuée, reste complexe. Nous avons constaté que les erreurs de transcription produisent des contre-sens qui viennent biaiser les résultats obtenus. Les passages erronés n'étant pas évidents à identifier, le TAL ne peut constituer, au moins sur la télé-réalité, qu'une aide pour le chercheur en sciences humaines. À charge pour l'humain de revenir vers le signal audio pour confirmer le résultat des approches de TAL appliquées sur les transcriptions. Je reviendrai sur des expériences de segmentation thématique de parole transcrite dans la section 5.5.

2.3 Hétérogénéité d'usages

Si les locuteurs d'une langue peuvent produire des contenus qui leurs sont propres, grâce à des moyens linguistiques, comme vu dans la section précédente, ils peuvent également introduire des variations dans leur production langagière, qui auront un impact sur la compréhension des textes produits, soit de manière inconsciente, parce que guidés par les possibilités technologiques, comme avec l'intégration d'émojis dans les messages produits depuis un téléphone portable (section 2.3.1), soit de manière consciente et délibérée, pour donner de la visibilité à sa production grâce aux hashtags (section 2.3.2), ou avec une dimension politique clairement affichée comme dans le cas du français inclusif (section 2.3.3).

2.3.1 Emojis (émoticônes, smileys)

Si le *smiley*, ce gros visage rond jaune et souriant, a été inventé en 1963 par le graphiste américain Harvey Ball [Daud and Ali, 2018], l'ajout d'émoticônes (séquence de caractères en ASCII) puis d'émojis—des marqueurs graphiques d'émotion qui ne sont pas limités aux émoticônes—dans sa production électronique a commencé avec les messages échangés par téléphone (SMS) et s'est poursuivi avec les micro-blogs. Des ajouts d'émojis dans les tables de caractères Unicode [22] sont par ailleurs régulièrement effectués [23], contribuant ainsi à l'adoption de ces éléments informatiques dans sa communication électronique jusqu'à demander de nouveaux ajouts.

À la faveur de la fonction de proposition de mots sur les téléphones portables, on observe une nouvelle tendance qui consiste à insérer des émojis dans sa communication électronique, non plus en complément des mots du message mais en remplacement de certains mots (figure 2.1). Sur cette figure, le message de gauche remplace le substantif *arbre* par l'émoji d'un sapin et le verbe *augmenter* par l'image d'une flèche orientée nord-est. L'émoji de la bouée vient renforcer le message de détresse « *Au secours* ». Les émojis utilisées sont suffisamment explicites pour qu'un locuteur comprenne le sens du message. Sur le message de droite, l'image d'un thermomètre renvoie à cet appareil de mesure et se trouve donc directement signifiant dans le contexte d'un été 2022 caniculaire, tandis que l'émoji de l'étoile remplace un substantif que je n'ai pas réussi à deviner (s'agit-il d'une marque ou d'un modèle de voiture, d'une pièce du domicile ?). Les autres émojis viennent compléter le contenu du message et renforcer l'idée véhiculée, de manière évidente pour le soleil (notions de chaleur et d'été), de manière plus subtile pour l'icône « shaka » ou « appel téléphonique de la main » (notion de relaxation). Le traitement des émojis doit tenir compte de leur signification, rendue plus complexe si la signification est restreinte à une communauté, et de leur intégration dans la chaîne écrite (en complément ou en remplacement d'un mot).

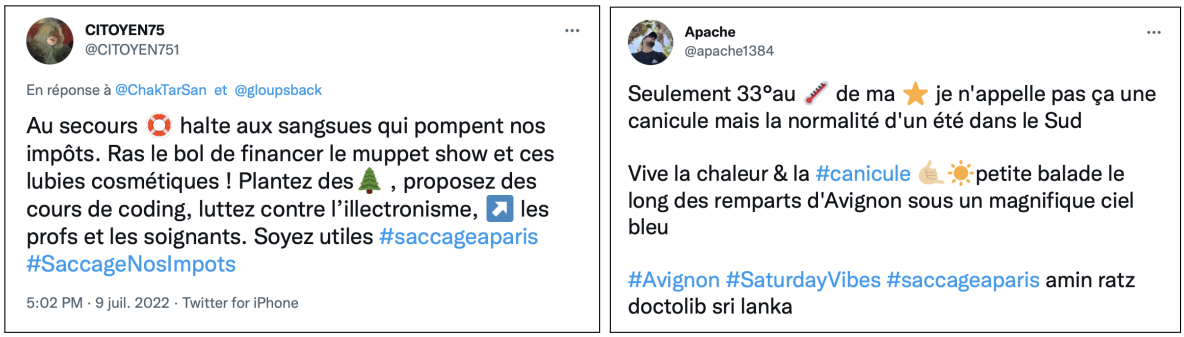

FIGURE 2.1 : Insertion d'émojis en complément ou en remplacement de mots de la langue

22. L'actuelle version d'Unicode 14.0, distribuée en septembre 2021, intègre 1404 codes représentant des émojis. Voir https://emojipedia.org/fr/unicode-14.0/

23. Parmi les ajouts envisagés dans Unicode 15.0, dont la sortie est prévue pour 2022, devraient figurer un visage tremblant, une main poussant vers la gauche, une main poussant vers la droite, une méduse, le gingembre, des maracas, etc. https://emojipedia.org/fr/unicode-15.0/

La principale difficulté dans le traitement de ce type de production repose sur le fait que le message écrit (texte en langue naturelle) se combine à des émojis, soit au format image (PNG), soit encodés en Unicode. Dans le cas de représentation sous forme de caractères Unicode, une correspondance peut être établie entre chaque icône et sa signification (modulo certaines icônes dont la signification est contextuelle), facilitant ainsi de futurs traitements du texte. L'intégration d'images rend plus complexe l'interprétation automatique du message et suppose l'utilisation de ressources externes (interrogation d'un moteur de recherche ou de l'encyclopédie des émojis [24]) ou de méthodes de traitement de l'image pour accéder au sens.

Les émojis constituent un marqueur fort pour la détection de polarité et l'analyse de sentiments [Guibon et al., 2016]. Des travaux ont porté sur la constitution de lexiques d'émojis pour l'analyse de sentiment [Ferández-Gavilanes et al., 2018]. Plus récemment, des modèles LSTM à base de plongements d'émojis ont été constitués pour améliorer l'analyse de sentiment [Chen et al., 2018], par exemple sur les tweets en chinois pendant la pandémie de Covid-19 [Liu et al., 2021].

2.3.2 Hashtags (mot-dièse, mot-clic)

Le développement du réseau social Twitter s'est accompagné d'une nouvelle pratique populaire reposant sur les *hashtags* (imparfaitement traduit « mot-dièse » en français et « mot-clic » au Québec), composé d'un ou plusieurs mots-clés sans espace et précédés du symbole croisillon (#), dans l'objectif de regrouper plusieurs messages d'une même thématique afin de faciliter le suivi des sujets d'intérêts [25]. Si les fils de discussion permettent de rassembler dans une même séquence le message initial et les réponses apportées, le hashtag est transversal et rassemble des messages provenant de discussions distinctes. Pour les recherches en TAL, ils facilitent la constitution de corpus, comme pour la campagne DEFT 2017 [Benamara et al., 2017] où les tweets constituant le corpus ont été identifiés grâce à deux types de hashtags : ceux qui permettent une catégorisation thématique du message (#DSK, #FIFA, etc.) et ceux qui constituent des marqueurs explicites du langage figuratif (#humour, #ironie, #joke, #sarcasme) [Karoui et al., 2017]. Dans la mesure où l'une des tâches consistait à identifier les tweets relevant du langage figuratif, ces hashtags ont été masqués dans les corpus distribués aux participants puisqu'ils ont servi à produire la référence.

Composition

Il existe plusieurs formes de hashtags selon leur composition : un seul mot (#canicule), plusieurs mots accolés sans espace (#BraderiedeLille, #pognondedingue) ou séparés par le caractère souligné (#nappes_phréatiques), combinaison de lettres et de chiffres (#Ligne2), et acronymes (#HCL pour *hydroxychloroquine*, #RATP). Sur le plan grammatical, il peuvent être assimilés à des syntagmes nominaux composés de substantifs ou de noms propres, et intégrés directement dans des phrases comme composants à part entière (exemples 15 et 16). Plus rarement, le hashtag renvoie à une phrase complète et grammaticalement correcte qui vient ponctuer le message (#mangezduriz, #jdcjdr pour « *je dis ça, je dis rien* ») ou se fonde sur un néologisme (#covid19, #escrologie) comme pour marquer un trait d'humour (combinaison du nom de ville La Clusaz avec l'acronyme ZAD de la « zone à défendre » sur l'exemple 16).

(15) Arrosage copieux ces prochains jours sur une grande partie du pays, sans caractère aggravant. Des pluies efficaces qui alimenteront les sols et les **#nappes_phréatiques** qui en ont encore bien besoin. (@lachainemeteo, 7 janvier 2023)

(16) Occupation du bois de la Colombière ! **#LaCluzad** s'organise ! La jeunesse est mobilisée soutenue par les habitants des montagnes et des vallées. Nous sommes allés les soutenir avec @binjamineurope ! Courage et #SauvonsBeauregard @Valerie_Paumier (@Fabienne-Grebert, 1er octobre 2022)

24. https://emojipedia.org
25. https://help.twitter.com/en/using-twitter/how-to-use-hashtags

Il est essentiel que les approches appliquées aux productions sur les réseaux sociaux tiennent compte des hashtags, en particulier s'ils s'insèrent dans la chaîne écrite, plus encore si des entités nommées sont présentes dans ces hashtags.

Usages

Il existe plusieurs usages possibles des hashtags, du point de vue de l'utilisateur[26] : catégoriser son message (exemple 17), ajouter du sens à son message (exemple 18 pour dénoncer un parti politique), marquer l'ironie ou le sarcasme (#jdcjdr ou #OhWait, sous-entendant une évidence, #Nawak), renseigner d'une émotion ou d'un état d'esprit (#GrosseFatigue), rechercher une information en interrogeant sa communauté professionnelle (#DocTocToc pour les docteurs en médecine sur l'exemple 19 par opposition à #ideTocToc pour les infirmiers diplômés d'État), ou encore associer son message à une marque.

(17) Après 3 semaines, complètement guéri et en pleine forme. Je me suis soigné par inhalation (confo + feuilles du manguier + Mutuzo + feuilles de menthe) et j'ai pris l' Arthemesia qui est très efficace. Prenez soin de vous et respectez les gestes barrières. **#COVID19** (@lemajestic1, 7 juillet 2021)

(18) Quand tu haïs tellement les riches que tu es prêt à financer des losers qui passent leur temps à traquer des jets privés... **#escrologie** (@LiberteNordiste, 28 août 2022)

(19) **#DocTocToc** A partir de quel âge surveiller particulièrement un enfant avec des ATCD d'otospongiose dans la famille ? Adolescence ? Puberté en particulier ? (@DocteurePATATE, 31 août 2022)

L'utilisation abondante en peu de temps d'un hashtag peut conduire ce hashtag à devenir une tendance (*trending topics*), renforçant la visibilité du sujet abordé, raison pour laquelle les médias incitent leur audience à réagir avec un hashtag qui mentionne le nom de l'émission, occasionnant en retour une publicité gratuite. Mais l'utilisation de hashtags peut donner lieu à des détournements opportunistes, par l'ajout de hashtags sans rapport avec le contenu développé dans les messages mais qui figurent en tendance, pour bénéficier d'une visibilité accrue (sur la figure 2.1 de droite, p. 24, le contributeur emploie le hashtag #saccageaparis et mentionne des références à l'actualité d'alors « Doctolib » et « Sri Lanka », alors que son message traite de la chaleur dans le sud et qu'il invite à se relaxer sur les remparts en Avignon).

Traitements

Plusieurs travaux ont porté sur les hashtags, soit pour recommander des hashtags à l'utilisateur [Mahajan et al., 2016, Ben-Lhachemi and Nfaoui, 2017], soit pour exploiter ces hashtags dans le cadre de tâches plus globales (repérage d'entités nommées, détection de polarité, classification, etc.). Plus spécifiquement sur la décomposition de hashtags, certains travaux s'appuient sur l'algorithme de Viterbi [Berardi et al., 2011], d'autres mobilisent des indices de casse typographique et la recherche de portions du hashtag en lexique [Brun and Roux, 2014, Belainine et al., 2016]. L'étude du contexte dans lequel apparaissent ces hashtags constitue également un domaine de recherche, en particulier pour améliorer le nettoyage des tweets [Henry et al., 2018]. Alors que cette tâche de décomposition pourrait passer pour un problème résolu, l'évolution des usages par les locuteurs de la langue vient ajouter de nouveaux cas à traiter, en particulier pour dénoncer le français inclusif (voir p. 28).

26. Je laisse de côté l'usage qui est fait par les services marketing d'entreprises et qui ont pour objectif d'augmenter l'engagement, regrouper des messages autour d'une marque sur plusieurs canaux de diffusion, contribuer à rendre viral un fait, et surveiller sa e-réputation.

2.3.3 Quand les utilisateurs·rices modifient la langue : le français inclusif

Le français inclusif est une variété du français standard, mise en avant pour témoigner d'une conscience de genre et d'identité [Alpheratz, 2018, Alpheratz, 2019] au regard de l'utilisation générique du masculin, actuellement employé pour rassembler un collectif composé de femmes et d'hommes (« *les lecteurs* »), ou dans les tournures impersonnelles (« *il fait beau* »). En raison du caractère performatif du langage [Austin, 1962], l'utilisation du masculin générique développe un habitus de pensée que les personnes promouvant le français inclusif cherchent à combattre. Il existe plusieurs procédés linguistiques d'atténuation ou de remplacement du masculin générique pour une production en français, tous rappelés par [Alpheratz, 2019].

La querelle des modernes et des anciens (le retour) Bien qu'attesté depuis plus d'un siècle dans les communications publiques sous la forme de coordination de gentilés au féminin et au masculin [27], le français inclusif s'est récemment illustré avec des polémiques et des prises de position autour de l'« écriture inclusive », mise en avant par le Haut Conseil à l'Égalité entre les femmes et les hommes [28] à l'origine d'un guide pratique [HCE, 2015]. Lors de la séance du 26 octobre 2017, l'Académie française a adopté à l'unanimité une déclaration considérant l'écriture inclusive comme une « *langue désunie* » qui crée « *une confusion qui confine à l'illisibilité* » et concluant que « *la langue française se trouve désormais en péril mortel* » [29]. Dans une lettre ouverte [30] du 7 mai 2021 cosignée par Hélène Carrère d'Encausse, Secrétaire perpétuel de l'Académie française, et Marc Lambron, Directeur en exercice de l'Académie française, les cosignataires dénoncent à la fois « *une injonction brutale, arbitraire et non concertée* » ainsi que le « *principe [d']une corrélation entre le genre des vocables et le sexe de leur référent* ». Suite à ces déclarations, une proposition de loi « *visant à sauvegarder la langue française et à réaffirmer la place fondamentale de l'Académie française* » a été enregistrée le 1er juin 2021 auprès de l'Assemblée nationale. À date de rédaction, cette loi est toujours à l'état de projet.

Une étude exploratoire Dans le cadre d'une étude exploratoire [Grouin, 2022], j'ai constitué un corpus d'une vingtaine de courts extraits de discours politiques français [31] publiés sur le site *Vie publique* [32] et de contenus publiés sur des sites gouvernementaux français tels que celui des données ouvertes publiques ou le *Journal officiel* [33], dans l'objectif d'une redistribution de ces ressources. Pour identifier les productions en français inclusif, j'ai listé empiriquement les locutions les plus fréquentes et cherché ces locutions dans les discours politiques. Parmi tous les procédés existants pour des productions en français inclusifs, deux émergent dans le discours politique : la coordination de formes féminines et masculines, ou doublets, à l'oral (p. 28), la combinaison de désinences féminines et masculines, également connue sous le nom d'« écriture inclusive », à l'écrit (p. 29). Dans les pages suivantes, je reviens sur les différents procédés identifiés dans le corpus politique. De manière complémentaire à ces différents procédés, réactiver l'ancienne règle d'accord en fonction de la proximité a été envisagée [Moreau, 2019], comme effectué dans le titre de la vidéo *les garçons et les filles sont belles* [Riban and Gerin, 2017].

27. Le quotidien *Le Drapeau* du 5 octobre 1910 reproduit la prise de parole de M. Dontenville qui commence par l'adresse « *Françaises, Français* » lors de la commémoration annuelle de la défense du Colonel Denfert-Rochereau lors du siège de Belfort (https://gallica.bnf.fr/ark:/12148/bpt6k40795430).
28. https://www.haut-conseil-egalite.gouv.fr/
29. https://www.academie-francaise.fr/actualites/declaration-de-lacademie-francaise-sur-lecriture-dite-inclusive
30. https://www.academie-francaise.fr/actualites/lettre-ouverte-sur-lecriture-inclusive
31. https://github.com/grouin/corpus-francais-inclusif, ce corpus sera régulièrement complété et annoté en fonction des données disponibles
32. https://www.vie-publique.fr/
33. https://www.data.gouv.fr/fr/ ; https://www.legifrance.gouv.fr/

Coordination de formes féminines et masculines (doublets)

Ce procédé est particulièrement prisé des politiques. Il est notamment resté dans les mémoires avec le discours « *Françaises, Français, aidez-moi !* » prononcé par le Général de Gaulle le 27 juin 1958, ou l'adresse « *travailleuses, travailleurs* » utilisée à chaque prise de parole publique par Arlette Laguiller depuis ses premiers discours de 1974. Au-delà de ces fameux exemples, on retrouve l'utilisation de ce procédé en tous temps, et de tout bord politique, de gauche (François Mitterrand, exemple 21) à droite (Jacques Chirac, exemple 23). La coordination vise essentiellement deux catégories de mots : les noms, essentiellement des gentilés (exemple 20), et des mots outils comme les pronoms démonstratifs (exemple 21) et les déterminants (exemples 22 et 23).

(20) **Martiniquaises, Martiniquais,** me voici donc à la Martinique où personne n'est jamais venu sans aimer à la fois cette terre et ce peuple. Voici donc **les Martiniquaises et les Martiniquais** rassemblés pour accueillir le President de la République qu'ils ont élu eux-mêmes démocratiquement. (Valéry Giscard-d'Estaing, 13 décembre 1974)

(21) Je remercie les enseignants, les professeurs, **celles et ceux** qui consacrent le meilleur de leur vie à la formation de la jeunesse de France. (François Mitterrand, 23 mai 1987)

(22) Merci à **toutes et tous** d'être nombreux aujourd'hui à Paris pour réaffirmer que le travail doit venir avant la Bourse. (Marc Blondel, 21 novembre 1998)

(23) Vous incarnez, en réalité, **toutes et tous**, l'Europe que nous souhaitons, c'est-à-dire l'Europe de l'union, l'Europe du progrès, l'Europe de l'ambition. (Jacques Chirac, 28 avril 2005)

Depuis la première élection d'Emmanuel Macron à la Présidence de la République en 2017, on observe une utilisation accrue, voire excessive, de ces coordinations par les membres du Gouvernement et plus largement par les politique du parti LREM. Ces coordinations de formes féminines et masculines ont également été adoptées massivement par les politiques de gauche. À l'heure actuelle, cet usage s'étend de l'extrême-gauche au centre, et devient un marqueur linguistique permettant d'identifier le bord politique d'un locuteur. En réaction, les hashtags #cellezéceux et #toutezétous ont été lancés par la droite, avec une connotation négative et un marqueur d'ironie (exemple 24). Ils constituent néanmoins une forme de hashtag inclusif.

(24) Fascinant de voir comment le « **cellezéceux** », « **toutezétous** », est passé en 5 ans d'un tic de langage purement macronien quelque peu ridicule et un tantinet exaspérant à LA nouvelle façon de parler. (@Valent1Pierre, 24 mai 2022)

On observe même une nominalisation du hashtag #cellezéceux pour désigner ses adversaires politiques ou plus simplement les électeurs d'Emmanuel Macron (exemples 25 et 26).

(25) Rétablir la sanction ? Soit ! Mais même régime pour tout le monde y compris et surtout pour les **#cellezéceux** politiciens qui devraient donner l'exemple et qui passent à travers les mailles ou ont des peines "aménagées" #cahuzac, #rugy #Ferrand #solere #lagarde... (@langlois_manu, 6 mai 2021)

(26) Quand les **cellezéceux** qui ne savent pas écrire en bon français nous donnent des leçons de féminisme grammatical, on a envie de leur asséner furieux : « *Est icelle dame menacée plus gravement par l'invasion misogyne, que par le masculin pas androgyne !* ». #GrandRemplacement #Woke (@FloraBerthelot, 23 juillet 2022)

Mais l'utilisation de ces formes féminines et masculines conduit parfois les locuteurs à faire des erreurs d'accord en genre, par hypercorrection de son discours pour s'assurer du caractère inclusif. L'hypercorrection désigne une « *réalisation grammaticale "fautive" due à l'application déplacée d'une règle imparfaitement maîtrisée* » [Arrivé et al., 1986]. Il en est ainsi de Sandrine Rousseau qui, le soir de son élection à la députation en juin 2022, remercie ses militants femmes et hommes pour « *la campagne que vous avez fait et faite* ». Qu'on soit femme ou homme, une seule formulation correcte en français, même inclusif : *la campagne que vous avez faite*.

Combinaison de désinences morphologiques féminines et masculines (« écriture inclusive »)

Le deuxième procédé consiste à combiner des désinences morphologiques masculine et féminine dans une même unité lexicale au moyen d'un point médian entre chaque flexion (exemple 27). Cette forme particulière d'écriture est qualifiée d'« écriture inclusive ».

(27) Une trace est la succession des inscriptions d'**un·e étudiant·e** à l'université en partant de son Bac jusqu'à sa dernière inscription. Une étape (de diplôme) est ce à quoi l'**étudiant·e** s'inscrit. Dans la version publique, présentée ici on masque l'effectif des cohortes de moins de 10 **étudiant·e·s**. (Visualisation des traces des étudiant·e·s de UP13, 6 mai 2017)

Plus spécifiquement pour la production de textes informatiques, l'absence de touche sur le clavier pour générer facilement le point médian conduit les utilisateurs à préférer des variantes telles que *chercheur.euse, lecteur.ice* (utilisation du point), *chercheur-euse, lecteur-ice* (utilisation du trait d'union), *chercheur'euse, lecteur'ice* (apostrophe, plus rare), et *chercheureuse, lecteurice* (absence de séparateur) ; ce dernier cas se rencontre sur les mots courts tels que *toustes*. Préalablement à la recommendation d'utiliser le point médian, les personnes qui voulaient inclure femmes et hommes dans leurs messages mettaient la désinence féminine entre parenthèses (exemple 28), mais cette solution a été abandonnée, pour « ne pas mettre les femmes entre parenthèses » [34], ou plus fondamentalement, parce que ce signe n'est pas « *le plus propre à signifier l'égalité dans un régime qui s'en réclame* » [Viennot, 2022].

(28) Nous voulons élargir à toute la société les possibilités d'accès aux formes les plus élaborées du savoir scientifique et permettre à **tout(e) étudiant(e)** d'aller au bout de ses possibilités, avec le souci permanent de la validation des parcours et des acquis. (Jean-Luc Mélenchon, 24 février 2012)

Pour éviter cette mise entre parenthèses des femmes, des tentatives alternatives d'utilisation de la barre oblique ont été réalisées comme en 2012 par Philippe Poutou (exemple 29).

(29) Les **transexuel/es** doivent pouvoir changer d'état-civil et/ou de numéro de Sécu, sans passer par un parcours psychiatrique, par une opération chirurgicale ou une stérilisation et les **intersexué/es** ne doivent pas être **mutilé/es** à la naissance pour les faire correspondre à un sexe ou un autre. [...] La prostitution touche des dizaines de milliers de personnes, dont plus de 20 000 **étudiant/es**. (Philippe Poutou, 11 avril 2012)

Non seulement l'écriture inclusive pose des problème à l'écrit, mais elle constitue également une difficulté réelle lors du passage à l'oral : la phrase « *Les doctorant·e·s sont venu·e·s nombreux·ses.* » serait prononcée par décomposition des éléments de flexion (« doctorant point e point s ») à l'image du domaine *.fr* (« point f r ») dans les adresses électroniques, conduisant à une perte de fluidité du discours. La prononciation de « *nombreux·euses* » avec une consonne sourde sifflante suivie d'une consonne sonore sifflante est impossible en français et engendre généralement une assimilation [nɔ̃brøgzøz] mais reste complexe.

Comme pour le précédent procédé de coordination de formes féminines et masculines, l'usage de l'écriture inclusive s'accompagne d'erreurs d'accord. En réaction à une actualité impliquant le Président, la député Raquel Garrido utilise la féminisation des fonctions et l'écriture inclusive (avec un trait d'union plutôt qu'un point médian) sur la fonction « *Président-e-s* », mais elle oublie de reporter ces différentes variantes de genre sur le participe passé « *chargé* » qui s'accorde avec le nom, ce qui devrait logiquement produire la forme « *chargé-e-s* » (exemple 30).

(30) La Première Ministre, qui conduit la politique de la Nation au titre de l'art. 20 de la Constitution, et les **président-e-s** de l'@AssembleeNat et du @Senat, qui sont **chargés**, par la Constitution, de la loi, doivent s'opposer à cet abus de pouvoir de la part du Président-monarque. (@RaquelGarridoFr, 29 août 2022)

34. https://www.liberation.fr/france/2017/09/27/pretes-a-utiliser-l-ecriture-inclusive_1598867/

Féminisation des noms de fonction

Ce procédé correspond à une préconisation du HCE. On retrouve des exemples de ce procédé dans la parole politique (exemple 30 sur « Première Ministre ») et dans les arrêtés parus au *Journal Officiel* (exemple 31) sans qu'ils ne soulèvent de commentaires.

(31) Par arrêté de la ministre de la défense en date du 26 septembre 2005, Mme Charlier (Dominique, Thérèse, Marthe), épouse Dagras, est nommée au grade de **lieutenante-colonelle**, en qualité d'**officière** recrutée au titre de l'article 29 du statut général des militaires, pour occuper un emploi de spécialiste des affaires politiques pour une durée de deux ans à compter du 1er octobre 2005. (Arrêté)

En revanche, l'exemple 32 est intéressant à plus d'un titre. En premier lieu, il met en lumière la forme « *autrice* », largement employée sur les ondes publiques, et qui fait grincer des dents, indépendamment de son genre et de son positionnement à l'endroit du processus de féminisation. Alors que les termes « *actrice* », « *directrice* » sont employés sans problème, celui d'« *autrice* », pourtant fondé sur le même processus dérivationnel (*directeur*→*directrice*, *auteur*→*autrice*) n'est pas encore totalement accepté. En second lieu, il reproduit les deux formes de féminisation du mot masculin « *auteur* » dans le même document [35], ce qui témoigne de l'absence de consensus encore actuel sur le choix de l'une ou l'autre des deux formes utilisées.

(32) Giorgia Marras, jeune **autrice** italienne [...] Giorgia Marras, illustratrice et **auteure** de bande dessinée, est née à Gênes en Italie, en 1988. (BD 2020)

Utilisation de termes collectifs ou épicènes

L'utilisation de termes collectifs ou épicènes *(personne, élève, individu, journaliste, membre, politique)* revient à employer des formes qui sont valables quel que soit le genre de la personne. C'est également le procédé le plus complexe à repérer en corpus car l'intention de l'auteur n'est pas évidente. C'est par la répétition de ce phénomène et l'absence d'utilisation des autres procédés (en particulier celui de coordination de formes féminines et masculines et celui de féminisation des noms de fonction) qu'il est possible d'identifier l'utilisation de ce procédé. Son identification automatique dans les discours politiques en vue de créer un corpus du français inclusif s'avère donc très complexe à réaliser.

Neutralisation

La neutralisation consiste à rendre neutre son discours, du point de vue du genre grammatical. En ce sens, il s'agit de recréer un genre neutre, qui « *s'oppose au masculin et au féminin par des marques formelles ou contextuelles* » [Mounin, 2004, p. 231]. Ce procédé repose sur une créativité morphologique fondée sur des constructions de type mot-valise ou en reprenant des morphèmes existants dans la langue, ou encore en puisant de nouvelles racines dans le substrat gréco-latin. Pour autant, ce procédé est actuellement totalement inutilisé des politiques et reste marginal, employés par les militants du français inclusif.

À l'heure actuelle, les constructions de type « mot-valise » ont principalement été mises en œuvre, notamment pour les pronoms personnels **iel* et **iels* (par combinaison des formes *il, elle* et *ils, elles*) [36] et dont l'intégration en novembre 2021 dans l'édition en ligne du *Robert* [37] a relancé la polémique sur le français inclusif, ou encore pour les pronoms démonstratifs **cellui* et **celleux* (par combinaison des formes *celle, celui* et *celles, ceux* ; la forme **ceuxelles* a également été proposée). La réactivation de formes du français médiéval (le pronom *el* ou *al* [Marchello-Nizia, 1989]) a

35. https://www.bd2020.culture.gouv.fr/media/BD-2020/cp-toute-la-france-dessine-giorgia-marras3
36. Les formes alternatives suivantes ont également été proposées, sans réel succès : **yel, *ael, *el, *ille, *ol, *ul.*
37. https://dictionnaire.lerobert.com/definition/iel

également été envisagé dans les tournures impersonnelles (**al fait beau*). Des créations ont également été réalisées de manière militante et à fin d'exemple [38], tel que le nom **frœur* pour englober *frères* et *sœurs*, ou **adelphité* en remplacement de *fraternité* qui réfère à la fratrie nécessairement masculine, sans que ces créations ne soient reprises à large échelle.

Une initiative intéressante à mes yeux, consiste à utiliser une graphie spécifique pour ce genre neutre, telle que la graphie « æ » dans des mots outils (**mæ, *lesquæls*) ou comme participe passé (*je suis *blessæ*, sans modification de la prononciation), et à employer des désinences existantes en français qui ont déjà cette valeur neutre, comme la désinence *-aire* dans *commissaire* ou *notaire* qui permet de créer aisément le néologisme **lectaire* sans que la compréhension n'en soit perturbée. Une approche complémentaire consiste à conserver les formes actuelles du masculin pour le neutre et à créer de nouvelles formes pour le masculin, permettant ainsi d'utiliser la forme courte pour le neutre et des formes plus longues pour le féminin et le masculin. On retrouve ce choix dans une nouvelle de Sylvie Lainé parue dans l'ouvrage collectif *Nos futurs* [Barre et al., 2020, p. 149] (exemple 33, la forme *gentil* est au neutre tandis que le néologisme **gentillï* constitue la nouvelle forme du masculin [39], à égalité en nombre de caractères avec la forme actuelle au féminin *gentille*) :

(33) Joséphine construit la phrase dans sa tête avant de la prononcer, à voix semi-basse.
— **Ellï** a l'air très **gentillï**, non ? Très **attentionnéï** ?
Elle a réussi à ramener un pâle sourire sur le visage fatigué de Lise.
— Tu n'es pas obligée d'être aussi formelle, tu sais. **Trevorï** aime bien qu'on le genre au masculin et qu'on en parle dans les formes exclusives, mais en public. En privé ça va. Ne t'embête pas. Tu peux utiliser le neutre. Oui, il est très **gentil**.

2.4 Conclusion

Dans ce chapitre, j'ai présenté les différents cas d'hétérogénéité que j'ai identifiés dans les corpus de production numérique sur lesquels j'ai travaillé ces dernières années. J'ai ainsi distingué l'hétérogénéité de contenu sur le plan linguistique, identifié lors du travail de thèse de Liyun sur les avis de touristes chinois à Paris, de l'hétérogénéité liée aux usages, imposés ou voulus par les utilisateurs, que j'ai observée dans mes travaux liés à la pharmacovigilance sur les réseaux sociaux, ou plus récemment à partir de transcriptions automatiques de la parole d'émissions radio et télévisées d'information et de divertissement.

Alors que le traitement automatique de la langue chinoise n'est déjà pas simple en soi, en raison de la concomitance de plusieurs systèmes d'écriture (sinogrammes traditionnels et simplifiés, pinyin) et du fait que les séquences de sinogrammes sont produites sans espace entre tokens, posant problème pour la tokénisation et pour toutes les tâches appliquées en aval (étiquetage en parties du discours, extraction d'information, détection de polarité, etc.), la thématique abordée dans le corpus nous a permis d'identifier des phénomènes supplémentaires venant complexifier le travail du TAL en chinois, notamment la problématique de translittération des mots étrangers (lieux, marques), sans qu'elle ne soit spécifique à la thématique du tourisme pour autant. Nous avons également constaté une hétérogénéité de styles liée au *code switching* et à l'utilisation de certaines figures de style propres aux réseaux sociaux. Le traitement des transcriptions automatiques de la parole constitue un champs d'application difficile, rendu encore plus complexe par la thématique traitée de la télé-réalité. J'ai pu constater que des choix de style individel utilisés dans ce type d'émission relèvent d'un enjeu de monétisation future de ses propres performances. Dans le chapitre 3, je présenterai le travail réalisé par Liyun autour de la fouille d'opinion et des inférences sur le corpus d'avis touristiques sur les réseaux sociaux.

38. Je reproduis ici les propositions issues de [Alpheratz, 2019]
39. Dans cette construction de nouvelles formes du masculin, Gabriel Illouz me faisait remarquer que le caractère « *ï* » constitue ici une représentation graphique symbolisant les organes sexuels masculins.

L'immédiateté des communications numériques sur les micro-blogs constitue un intérêt pour plusieurs champs d'application tel que la pharmacovigilance, sur laquelle je reviendrai dans le chapitre 4, mais la communication sur ces micro-blogs s'accompagne de pratiques qui leur sont propres, tels l'ajout d'émojis comme marqueurs graphiques d'émotions, ou de hashtags pour indexer et lier plusieurs messages. Si l'utilisation de hashtags comme constituants d'une phrase (ou pour le dire autrement, la transformation d'un mot ou d'une expression polylexicale de la phrase en hashtag) est un phénomène qui existe depuis que les hashtags sont utilisés, on observe un phénomène similaire avec les émojis, sous l'effet des modules de production de texte sur téléphone portable, qui proposent désormais de remplacer certains mots de la phrase par l'émoji correspondant. Ces nouveaux usages, parfois imposés par la technologie contre la volonté des utilisateurs, appellent la modification des approches de TAL pour accéder au sens contenu dans ces différentes productions langagières.

Enfin, et sans que cela ne soit spécifique aux réseaux sociaux, les locuteurs modifient parfois la langue dans un objectif politique ou militant. C'est notamment le cas du français inclusif sur lequel j'ai commencé à travailler, en tant qu'objet d'étude, et qui offre des perspectives de recherche intéressantes au point de vue TAL, notamment par la combinaison des différentes propriétés observées dans ce chapitre (en particulier la nominalisation du hashtag inclusif #cellezéceux dans un objectif de contestation et de dénonciation de la parole politique).

Chapitre 3

Les non-dits des utilisateurs

Sommaire

3.1 Introduction

3.1.1 Énoncé et discours

L'*énoncé* s'inscrit, pour Benveniste, dans un contexte d'*énonciation*, défini comme une « *mise en fonctionnement de la langue par un acte individuel d'utilisation* » [Benveniste, 1980, p. 80]. L'énonciation est une « *réalisation individuelle* ». Benveniste distingue la *langue* en tant que possibilité, avant énonciation, de la langue instanciée en discours avec « *appropriation* » par le locuteur, après énonciation, à l'image de la distinction opérée par Saussure entre *langue*, qui relève de l'universel, et *parole*, qui relève de l'individuel [Saussure, 1916]. Cette distinction rejoint également l'opposition formulée par Chomsky entre la *phrase* qui relève de la compétence et l'*énoncé* qui relève de la performance [Chomsky, 1957]. La phrase serait donc générique, au niveau de la langue, alors que l'énoncé serait instancié par l'individu. L'énonciation mobilise des *embrayeurs*, notion proposée par Jakobson pour désigner des marqueurs linguistiques qui vont inscrire l'énoncé dans le temps, l'espace, et la personne [Jakobson, 1963, Carlotti, 2011].

Plus complexe est la notion de *discours* [Maingueneau, 1979]. Pour Saussure, cette notion est une variante de celle de parole. Il s'agit d'une instanciation individuelle. Pour Guespin, un texte « *du point de vue de sa structuration "en langue" en fait un énoncé* » alors qu'« *une étude linguistique des conditions de production de ce texte en fera un discours* » [Guespin, 1971]. On retrouve ici l'opposition saussurienne entre langue et parole. Pour Benveniste, l'énonciation est une « *instance de discours qui émane d'un locuteur, forme sonore qui atteint un auditeur et qui suscite une autre énonciation en retour* » [Benveniste, 1980, p. 81–82]. Le discours se compose d'un ensemble d'énonciations entre interlocuteurs. Pour Charaudeau, un discours se constitue dans l'« *acte de langage* » et se rapporte à « *un ensemble cohérent de savoirs partagés* » [Charaudeau, 1984]. Le discours apparaît ici comme l'inclusion d'un énoncé dans un contexte de communication, où les interlocuteurs partagent des savoirs.

À titre de synthèse, nous retiendrons que la langue est un système de compétences universelles, partagé par une communauté de locuteurs. L'instanciation individuelle de la langue, au moyen d'embrayeurs, est une performance qui produit un énoncé. Un ensemble d'énoncés dans le cadre d'une communication entre interlocuteurs produit un discours, qui relève d'une *compositionnalité* [Moeschler, 2017].

3.1.2 Inférences

La compréhension entre individus repose souvent sur de l'implicite en même temps que de la subjectivité, que chaque locuteur saura interpréter grâce à son expérience et ses connaissances sur le monde. Benveniste rappelle l'existence de conventions de politesse qui imposent « *l'emploi de périphrases ou de formes spéciales entre certains groupes d'individus, pour remplacer les références personnelles directes* » dans les sociétés asiatiques [Benveniste, 1976, p. 261]. Cette observation porte sur l'omission des pronoms personnels, mais elle s'applique plus largement encore, notamment concernant l'expression des opinions. Ces dimensions culturelle et sociétale identifiées dans les productions langagières renvoient à l'ordre herméneutique présenté en introduction (voir section 1.2). Dans cet ordre, Rastier rappelle le lien qui existe entre production langagière et interprétation de ces productions, par le biais des connaissances culturelles et sociétales [Rastier, 1996]. Nous avons exploité ce lien au travers des inférences pour améliorer la fouille d'opinion en complément de la recherche habituelle de termes porteurs d'opinion, sentiment, émotion (OSE).

L'inférence est une opération logique qui admet pour vraie une proposition en raison de sa relation avec d'autres propositions, qualifiées de prémisses, qui auront été préalablement considérées comme vraies. Le raisonnement par déduction correspond au *syllogisme* formulé par Aristote (384–322 av. J.-C.) dans son *Organon* et explicité au moyen d'une prémisse principale « *Tous les hommes sont mortels* » et d'une prémisse secondaire « *Socrate est un homme* », où on arrive à la conclusion que « *Socrate est mortel* ». Plus récemment, sans pour autant être contemporain, Peirce (1839–1914) a opposé les concepts de *déduction*, *induction*, et *hypothèse* dans ses travaux de sémiotique, de manière à classer les inférences [Peirce, 1878].

— déduction : soit la règle générale « *Tous les haricots du sac sont blancs* » et le cas particulier « *Ces haricots proviennent de ce sac* », on conclut (résultat) que « *Ces haricots sont blancs* ». La déduction correspond au syllogisme d'Aristote. Il s'agit d'un raisonnement explicatoire. En outre, Peirce considère que la déduction permet la prédiction d'effets.

— induction : soit le cas particulier « *Ces haricots proviennent de ce sac* » et l'observation « *Ces haricots sont blancs* », on induit une règle générale « *Tous les haricots du sac sont blancs* » [1]. L'induction relève d'un processus de généralisation à partir d'un cas et d'un résultat constaté. Ce raisonnement produit une règle qui relève de la vraisemblance.

— hypothèse : soit la règle générale « *Tous les haricots du sac sont blancs* » et l'observation « *Ces haricots sont blancs* », on émet l'hypothèse que « *Ces haricots proviennent de ce sac* » [2] L'hypothèse relève d'une supposition effectuée à partir d'une règle générale et d'un constat.

Sur cette base définitoire, Peirce a proposé une classification des inférences (figure 3.1) en distinguant les inférences déductives (ou analytiques) des inférences synthétiques qui reposent sur l'induction et l'hypothèse.

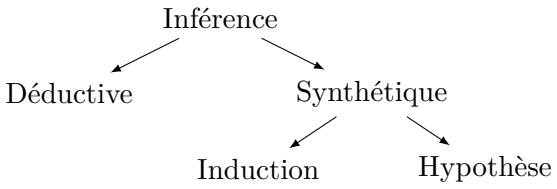

FIGURE 3.1 : Classification des inférences selon Peirce

On pourrait y voir une reprise de la distinction formulée par Kant (1724–1804) dans sa *Critique de la raison pure* qui distinguait les raisonnements nécessaires ou explicatifs fondés sur la

1. Appliqué à l'exemple d'Aristote, l'induction consisterait, à partir du cas « *Socrate est un homme* » et du résultat « *Socrate est mortel* », à induire la règle « *Tous les hommes sont mortels* ».

2. Appliqué à l'exemple d'Aristote, à partir de la règle « *Tous les hommes sont mortels* » et du résultat « *Socrate est mortel* », on émet l'hypothèse que « *Socrate est un homme* ».

déduction, des raisonnements ampliatifs fondés sur l'induction [Kant, 1781]. Certains travaux ont cependant mis en avant que la pensée de Peirce et le paradigme qu'il a formulé sont précisément en rupture vis à vis des travaux d'Aristote, Descartes et Kant [Deledalle, 1994].

Dans le cas de la déduction, la conclusion reste vraie tant que les prémisses le sont également. Si la conclusion ne correspond pas à ce qui est attendu, ce que Peirce appelle un « fait surprenant », c'est que l'hypothèse d'origine doit être revue puisqu'elle ne permet pas d'expliquer ce fait. Elle doit l'être dans la mesure où les processus de déduction (prédiction) et d'induction (généralisation) qui étaient valables et confinaient à cette hypothèse son caractère vraisemblable ne sont désormais plus corrects. Ce fait surprenant constitue le point de départ de l'*abduction*, qui suppose un retour en arrière pour formuler une nouvelle hypothèse, d'où le concept de *rétroduction* qui a parfois été préféré [3], même si certains travaux s'appuyant sur l'étymologie des préfixes considèrent que ces deux concepts ne sont pas interchangeables [Chiasson, 2005]. Si la distinction entre déduction et induction fait consensus, celle d'abduction est davantage discutée [Roudaut, 2017], y compris par Peirce lui-même qui y aura consacré plusieurs années de recherche en proposant des définitions parfois contradictoires [Dumez, 2012]. Pour autant, il semble acté que la logique de Peirce peut s'organiser entre abduction *(Primséité)* et déduction *(Secondéité)* pour expliquer des faits de manière empirique mais pas pour apporter de nouvelles connaissances, et induction *(Tiercéité)* pour enrichir des connaissances en générant de nouvelles hypothèses [Deledalle, 1994, Chong Ho, 1994]. J'ai mis en pratique ces théories à l'occasion de la thèse de Liyun Yan [Yan, 2021], dans un objectif d'apporter des connaissances linguistiques supplémentaires pour la fouille d'opinion.

3.2 Proposition d'une classification des inférences

Le travail de thèse réalisé par Liyun Yan autour des inférences repose sur les opinions véhiculées dans des commentaires d'internautes chinois. Le corpus sur lequel nous avons travaillé se compose de 23 000 avis de touristes sinophones (Chine continentale, Hong Kong et Taïwan) en visite à Paris (intra-muros uniquement) déposés entre janvier 2016 et décembre 2017, donc en dehors d'une période pouvant biaiser les analyses telle que la période Covid-19, sur deux sites de réservation en ligne, un site international [4] et un site chinois [5], en identifiant les hôtels communs aux deux plateformes. D'autres sites internationaux ont été identifiés mais exclus, en raison de la présence de messages en chinois provenant d'une traduction automatique. Ces traductions ne reflètent pas l'avis de touristes chinois mais correspondent potentiellement à un avis rédigé en français et traduit automatiquement, dans lequel on ne retrouve aucun élément de tradition culturelle chinoise. Lors de la constitution du corpus, nous avons émis l'hypothèse que les avis déposés sur le site chinois seraient plus représentatifs de touristes de culture chinoise, formulant des attentes précises par rapport à leurs habitudes de vie, alors que les avis sur le site international pourraient refléter des profils de voyageurs plus ouverts sur le monde. En d'autres termes, la répartition des types d'inférences sera différente en raison de l'audience distincte sur les deux sites. Ce corpus a fait l'objet d'un découpage automatique en mots, pour un total de 2045 mots, puis d'un étiquetage en parties du discours au moyen de l'outil Jieba [6] et d'une liste des principaux sites parisiens produite par Liyun.

Dans l'exemple 1, nous comprenons qu'il s'agit d'une opinion positive puisque le ménage est réalisé à un moment qui ne dérange pas les occupants. Il n'est pas nécessaire de mobiliser des connaissances particulières pour comprendre le contenu du message.

(1) 房间整理及时. (fángjiān zhēnglǐ jíshí). *Chambre nettoyée au bon moment.*

3. S'il s'agit de regarder en arrière, le préfixe latin *retro-* favorise le concept de rétroduction.
4. https://www.booking.com/
5. http://www.mafengwo.cn/
6. https://github.com/fxsjy/jieba ; l'outil offre la possibilité de prendre en entrée un dictionnaire personnel pour affiner les résultats.

Dans les exemples 2, 3 et 4, des connaissances culturelles sont nécessaires pour interpréter correctement l'opinion véhiculée. Si la négation permet déjà de supposer un avis négatif, l'implicite culturel confirme cette valence négative dans la mesure où la bouilloire constitue un équipement indispensable pour un touriste asiatique souhaitant pouvoir faire du thé à tout moment (exemple 2). De manière similaire, le locuteur doit savoir que les Galeries Lafayette sont des galeries commerciales prisées des touristes pour comprendre que la proximité de ces grands magasins constitue un avantage dans la localisation de l'hôtel (exemple 3).

(2) 没有烧水壶. (méiyōu shāo shuǐhú). *Pas de bouilloire.*

(3) 离老佛爷近. (lí lǎo fóyé jìn). *Proche des Galeries Lafayette.*

Dans l'exemple 4, une différence culturelle fait jour concernant la taille des cabines d'ascenseur. Alors qu'un européen pourrait apprécier l'héritage architectural et industriel du fer forgé de la grille fermant la porte palière, ou simplement l'existence même d'un ascenseur dans un immeuble haussmannien, un touriste asiatique considèrera d'abord d'un regard critique l'espace réduit qui ne lui permet pas d'y faire entrer facilement des bagages généralement volumineux.

(4) 电梯迷你到不刻意找你都发现不了。(diàntī mínī dào bù kèyì zhǎo nī dōu fāxiàn bùliǎo). *L'ascenseur est si petit que vous pouvez ne pas le voir si vous ne le cherchez pas délibérément.*

Dans un premier temps, nous avons exploré les différents types d'inférences qui existent en nous fondant notamment sur les travaux théoriques de Peirce, et en nous inspirant de travaux similaires pour de la fouille d'opinion [Doucy and Massoussi, 2012]. Sur cette base et dans un objectif d'identification de la polarité exprimée dans des messages à des fins de fouille d'opinion, Liyun a proposé une typologie des inférences fondée sur trois principaux niveaux d'analyse : le type de réalisation sémantique (section 3.2.1), la modalité de réalisation (section 3.2.2), et le mode de production (section 3.2.3). Ce travail a donné lieu à une première publication, dans la conférence jeunes chercheurs RECITAL [Yan, 2018].

3.2.1 Réalisation sémantique

La réalisation sémantique repose sur le type de raisonnement qu'un locuteur doit effectuer pour accéder au sens exprimé par l'inférence. Cette réalisation peut-être *logique* si elle implique un raisonnement formel ou qu'elle repose sur une interprétation littérale du texte (exemple 5), *pragmatique* si elle nécessite que le locuteur fasse appel à des connaissances autres que purement linguistiques telles que des connaissances culturelles (exemple 6), ou *lexicale* lorsqu'il s'agit de termes qui, par essence, peuvent s'instancier en dehors de tout cadre énonciatif (exemple 7). Étant donné ce corpus d'avis touristiques, les inférences lexicales concernent souvent des noms propres référant à des monuments touristiques (exemple 8) ou de grandes marques d'hôtel, de magasin, ou de maroquinerie *(George V, Printemps, Versace)*.

(5) 前台只有一个人，非常忙碌。每次都要排队等待 (qiántái zhǐyǒu yīgè rén, fēicháng mánglù. měi cì dōu yào páiduì děngdài) *Il n'y a qu'une seule personne à la réception et elle est très occupée. Il faut toujours faire la queue*

(6) 电梯小 (diàntī xiǎo) *Petit ascenseur*

(7) 蟑螂 (zhāngláng) *cafard*

(8) 埃菲尔铁塔 (āifēi'ěr tiětǎ) *Tour Eiffel*

Ainsi, à la lecture de l'exemple 5, on comprend que l'avis est négatif sans avoir besoin d'autres connaissances que celles requises pour lire un texte. Inversement, dans l'exemple 6, il est nécessaire d'avoir une connaissance de la culture chinoise pour comprendre qu'un petit ascenseur ne sera pas pratique pour monter des bagages volumineux, ce qui conduit à un avis négatif sur cet équipement. Sur les exemples 7 et 8, les expressions « cafard » et « Tour Eiffel » impliquent, par essence, une

opinion négative pour la première[7] et positive pour la seconde, dans le contexte de tourisme à Paris (qui peut toutefois s'inverser en discours, si le locuteur exprime la fermeture de la Tour Eiffel, ou, situation plus rare, si un entomologiste cherchant des cafards est satisfait d'en trouver).

3.2.2 Modalité de réalisation

La modalité de réalisation désigne le processus mental mis en œuvre par le locuteur pour accéder au sens, parmi les trois opérations définies par Peirce : déduction, induction, et rétroduction. Nous reprenons la notion de déduction comme processus logique s'appuyant sur une prémisse pour aboutir à une conclusion, ce qui conduit à une *inférence immédiate* (exemple 9). De même pour l'induction, envisagée comme processus de généralisation, mais qui conduit à une *inférence médiate* (exemple 10). Concernant la rétroduction, nous la définissons comme une inférence qui fait appel à des connaissances antérieures exprimées dans la phrase ou le discours (exemple 11).

(9) 前台服务人员一直是笑容满面。(qiántái fúwù rényuán yīzhí shì xiàoróng mǎnmiàn) *Le personnel de la réception est toujours souriant*

(10) 周围有很多餐馆，也有家乐福超市。(zhōu wéi yǒu hěnduō cānguǎn, yěyǒu jiālèfú chāoshì) *Il y a de nombreux restaurants aux alentours, il y a aussi un supermarché Carrefour.*

(11) 前台非常热情，得知我丢失手机帮助我打电话联系国内，让我使用电脑等等，还给我一个人分了个四人间。(qiántái fēicháng rèqíng, dé zhī wǒ diūshī shǒujī bāngzhù wǒ dǎ diànhuà liánxì guónèi, ràng wǒ shǐyòng diànnǎo děng děng, hái gěi wǒ yīgè rén fēnle gè sì rénjiān) *La réception est très serviable. Quand ils ont appris que j'avais perdu mon téléphone portable, ils m'ont permis d'appeler chez moi, m'ont laissé utiliser l'ordinateur, etc., et m'ont même donné une chambre quadruple.*

On déduit immédiatement de l'exemple 9 une opinion positive alors que dans l'exemple 10, on induit que cette situation est positive parce qu'il s'agit de touristes qui auront envie d'aller au restaurant et que la présence de supermarchés dans les environs de l'hôtel est positive. Dans l'exemple 11, la serviabilité de la réception s'explique par l'ensemble des actions détaillées dans la suite du message en faveur de l'infortunée personne, ce qui conduit à un avis positif sur l'hôtel.

3.2.3 Mode de production

Enfin, le mode de production renvoie au cadre dans lequel l'émetteur du message a produit l'inférence, parmi deux modes : *énonciatif*, en considérant qu'un énoncé est une instanciation individuelle actualisée en contexte (exemple 12) ou *discursif*, le discours étant ici considéré comme un enchaînement cohérent d'une suite de syntagmes [Harris, 1952] (exemple 13).

(12) 离地铁站很近 (lí dìtiě zhàn hěn jìn) *très proche du métro*

(13) 第二天一次性拖鞋就穿坏了，打扫的服务员都不会及时更换，我同行的朋友去前台要，结果一双都没有找到，这些- 应该配备的。(dì èr tiān yīcì xìng tuōxié jiù chuān huàile, dǎsǎo de fúwùyuán dōu bù huì jíshì gēnghuàn, wō tóngxíng de péngyǒu qù qiántái yào, jiéguǒ yīshuāng dōu méiyǒu zhǎodào, zhèxiē-yīnggāi pèibèi de) *Le lendemain, les pantoufles jetables étaient usées, et le personnel de nettoyage n'a pas voulu les remplacer à ce moment. Mon ami est allé à la réception pour en demander mais ils n'ont pas trouvé une seule paire. Ils devraient en être équipés.*

L'exemple 12 témoigne d'une localisation d'hôtel à proximité immédiate d'une station de métro, ce qui est positif. Dans l'exemple 13, on comprend que l'ensemble de la séquence est perçue négativement par le touriste, par l'enchaînement de situations multiples (usure de la paire, absence de remplacement par le personnel de ménage, indisponibilité de nouvelles paire à la réception).

7. Que le terme « cafard » soit pris au sens propre « *j'ai vu des cafards* » ou au sens figuré « *j'ai le cafard* », l'opinion exprimée par cette inférence lexicale reste négative dans les deux cas pour la plupart des personnes.

3.2.4 Représentation en corpus

Un sous-corpus composé de 69 commentaires rédigés en chinois simplifié a été manuellement annoté par Liyun et Yihong, une étudiante en licence de linguistique informatique. Ce premier travail d'annotation a été envisagé comme une preuve de concept, en particulier sur la possibilité de représenter les différents types d'inférence. Les avis concernent trois hôtels parisiens, sélectionnés pour leur diversité géographique et de standing [8], en observant une distribution équilibrée des avis selon le site d'origine [9]. Nous présentons sur les figures 3.2 et 3.3 des exemples d'annotations humaines des inférences dans deux commentaires du même hôtel.

Le premier message se compose de six propositions : *la chambre est très petite* (房间非常小), *mais elle est propre* (但是还算整洁), *la localisation géographique est très bien* (地理位置非常棒), *proche de la sortie du métro* (出门地铁站), *proche du grand magasin du Printemps* (近春天百货), et *le restaurant en bas de l'hôtel est célèbre et délicieux* (楼下餐馆有名并且好吃).

Les deux propositions « *proche de la sortie du métro* » et « *proche du grand magasin du Printemps* » ont toutes deux été annotées comme des inférences pragmatiques déductives discursives, avec un élément lexical ayant, par essence, une valence positive (en l'occurrence, 地铁站 « *station de métro* » et 春天百货 « *grand magasin du Printemps* »). Les autres propositions n'ont pas été considérées comme instanciées dans un cadre inférentiel.

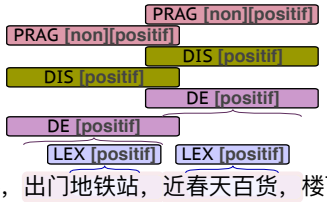

FIGURE 3.2 : Annotation humaine d'inférences réalisées au niveau pragmatique avec un élément lexical (LEX), de modalité déductive (DE), dans un mode discursif (DIS), pour des opinions positives (« proche de la sortie du métro » et « proche du grand magasin du Printemps »)

Dans le deuxième message, nous relevons trois propositions : *la salle de bain est très petite* (卫生间的沐浴间非常小), *il faut tourner avec précaution* (要小心翼翼的转身), et *la chambre pour 4 personnes est très spacieuse* (4 人间卧室非常宽敞). Les deux propositions portant sur la salle de bain et le type de mouvement à y effectuer sont rassemblées dans une inférence logique énonciative et constituent, fort logiquement, une opinion négative, d'autant plus que la chambre est qualifiée de spacieuse. Cette dernière proposition ne semble cependant pas contenir d'inférence.

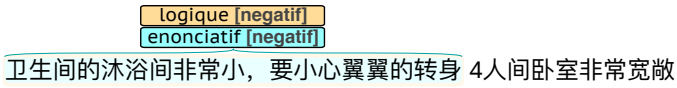

FIGURE 3.3 : Annotation humaine d'inférence réalisée au niveau logique, dans un mode énonciatif, pour une opinion négative (« la salle de bain est très petite, il faut tourner avec précaution »)

En raison de la difficulté rencontrée par les deux annotatrices pour annoter les modalités de réalisation (déduction, induction, et rétroduction), en partie liée à la complexité de la pensée de Peirce [Chong Ho, 1994, Dumez, 2012], et par les probables difficultés que rencontreraient des outils du TAL pour identifier automatiquement ces phénomènes, nous avons fait le choix de ne

8. Un hôtel 3 étoiles, rue Saint-Lazare (8ème), à proximité des grands magasins du boulevard Haussmann, un hôtel 4 étoiles proche de la Tour-Eiffel, rue du Général de Larminat (15ème), et un établissement 5 étoiles de type appart'hôtel, avenue des Champs-Élysées (8ème).

9. Soit 20 messages pour l'hôtel 3 étoiles (pour moitié de Booking, l'autre moitié de TripAdvisor), 29 messages pour l'hôtel 4 étoiles (à raison d'un tiers par site entre Booking, Mafengwo, et TripAdvisor), et 20 messages pour l'hôtel 5 étoiles (pour moitié de Booking et de Mafengwo).

pas poursuivre l'effort d'annotation sur ces informations dans la suite des travaux de thèse de Liyun. Ce travail exploratoire nous aura permis de prendre conscience de la difficulté de la tâche, ce qui suggère de revoir ou de simplifier le travail d'annotation des modalités, mais également de constater la richesse des informations linguistiques qui existent dans ce corpus.

3.3 Application à la fouille d'opinion

3.3.1 Annotation manuelle

Un corpus plus conséquent de 1391 avis touristiques (correspondant à 499 hôtels) composés d'au-moins huit mots par avis, a été constitué par échantillonnage (localisation géographique, qualité touristique en nombre d'étoiles, et distribution équilibrée entre messages positifs et négatifs fondée sur le résultat d'une analyse textométrique produite par l'outil TXM [Heiden et al., 2010, Heiden, 2010]). L'annotation a consisté à identifier les passages contenant une opinion, et pour ces passages, à annoter les différents niveaux d'analyse suivants lorsque cela s'est révélé possible :

- *opinion* en termes d'opinion et de cible, avec annotation de la relation entre l'opinion et la cible pour déterminer si le lien est explicite ou implicite. Pour les cibles, la thématique visée est précisée parmi dix thèmes en lien avec l'hôtellerie dont : chambre, localisation, service
- *valence* des portions annotées parmi quatre classes : positif, neutre, négatif, ou inconnu
- *présence* d'une inférence dans les portions annotées parmi trois classes : absence d'inférence, présence d'une inférence, et incertitude sur la présence d'une inférence.

  Pour les inférences qui auront été jugées présentes uniquement, annotations complémentaires de deux types d'information, d'après les travaux de classification des inférences précédemment réalisés (voir section 3.2) :

  - *réalisation sémantique* de l'inférence parmi trois classes : logique, pragmatique, lexicale
  - *mode de production* de l'inférence parmi deux classes : discursif, énonciatif
- et vérification des parties du discours proposées par Jieba, parmi onze classes [10] : adjectif, adverbe, conjonction, construction (préposition « de »), mot étranger (*code switching* et emprunts), négation, particule, préposition, pronom, substantif, verbe. Pour trois classes, une spécification est possible : adjectif qualificatif, adverbe modifieur, nom de lieu.

Ce travail a été réalisé par Liyun Yan, aidée de trois stagiaires [11] (Mei Gan, Danni E, Yuning), dans la perspective de démontrer l'intérêt d'étudier les inférences pour la détection automatique de la polarité [Yan et al., 2020]. Les accords inter-annoteurs calculés au moyen de la F-mesure varient de 0,938 à 0,963 pour le simple repérage d'inférence (présence, absence, incertitude), et diminuent de 0,651 à 0,705 pour l'identification des types d'inférence (réalisation sémantique et mode de production). J'estime ces résultats corrects au regard de la complexité à identifier ce type d'information linguistique, et utilisables pour de futures expériences d'apprentissage. Le repérage des inférences lexicales a été le plus simple, tandis que l'identification du mode discursif a posé le plus de difficultés aux annotatrices. L'information de polarité ne semble pas aussi évidente qu'on pourrait le supposer, avec des accords qui varient de 0,817 à 0,845. Il en est de même pour l'information du sujet traité dans les portions inférentielles, avec des valeurs comprises entre 0,706 et 0,745. À l'issue de ce travail, le corpus se compose de 7972 portions annotées [12].

10. La répartition des classes dans le corpus manuellement annoté est la suivante : 553 substantifs, 417 adverbes, 342 verbes, 336 adjectifs, 93 particules, 72 pronoms, 71 conjonctions, 57 prépositions, 35 négations, 16 prépositions de construction, et 7 mots étrangers *(cooking, roissybus, view, wifi, wi-fi)*.

11. L'INaLCO a financé trois stages de courte durée (4h/jour pendant un mois) pour des étudiantes chinoises en licence. À cette occasion, Liyun aura fait l'apprentissage de l'encadrement d'une équipe et de la distribution des annotations en fonction des qualités de chaque annotatrice (qualité des annotations réalisées et rapidité d'exécution).

12. Corpus annoté de 1391 avis (499 hôtels) : https://github.com/liyunyan/ChineseHotelReviewAnnotation

3.3.2 Représentation en corpus

Nous présentons sur les figures 3.4 et 3.5 des exemples d'annotations humaines des opinions et cibles, des relations entre opinion et cible, et de vérification des parties du discours dans deux commentaires du même hôtel (mêmes exemples que ceux des figures 3.2 et 3.3 pour faciliter la comparaison ; les annotations d'inférences n'ont pas été reportées sur ces exemples pour éviter de surcharger la représentation des textes).

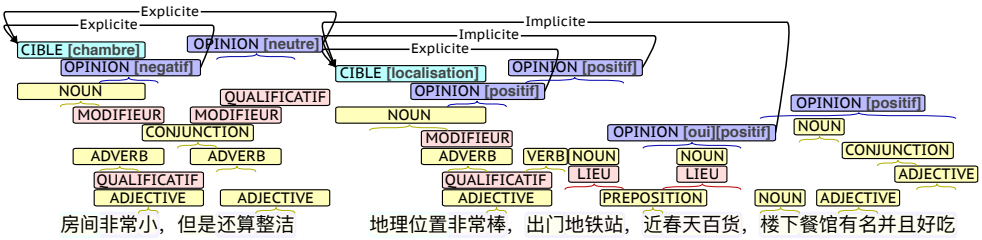

FIGURE 3.4 : Annotations humaines d'opinions négative (« très petit ») et neutre (« assez propre ») explicitement liées à la chambre, suivies d'opinions positives (« très bien », « sortie du métro », « proche du grand magasin du Printemps ») explicitement et implicitement liées à la localisation, et d'une opinion positive qui n'est liée à aucune cible (« le restaurant en bas est célèbre et délicieux »)

Dans ce premier message, six opinions sont exprimées, à raison d'une opinion négative (« *très petit* ») explicitement liée à la cible « *chambre* » sur la thématique de la chambre d'hôtel, d'une opinion neutre (« *assez propre* ») également liée de manière explicite à cette même cible, suivies de quatre opinions positives (« *très bien* » et « *sortie du métro* » explicitement reliées à la cible « *localisation géographique* » sur la thématique de la localisation, « *proche du grand magasin du Printemps* » également liée à cette même cible de manière implicite et relevant d'une inférence, et « *le restaurant en bas est célèbre et délicieux* » qui n'est reliée à aucune cible).

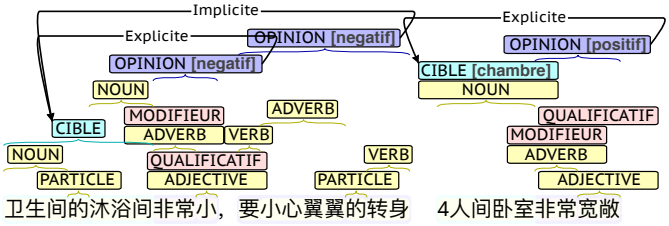

FIGURE 3.5 : Annotations humaines d'opinions négative (« très petit »), neutre (« tourner avec précaution »), et positive (« très spacieux »), en lien avec deux cibles (« salle de bain » et « chambre »), sur la thématique de la chambre d'hôtel

Dans ce deuxième message, trois opinions sont exprimées : la première est négative (« *très petit* ») et explicitement liée à la cible « *salle de bain* » ; la deuxième est neutre (« *il faut tourner avec précaution* ») et implicitement liée à la cible précédente en ce sens que la proposition se rapporte au sujet exprimé dans la première proposition ; la dernière opinion est positive (« *très spacieux* ») et explicitement liée à la cible « *chambre pour 4 personnes* ».

Nous observons dans ces exemples que toutes les opinions se composent d'adjectifs qualificatifs et d'adverbes modifieurs qui servent de marqueur d'intensité. Les cibles concernent uniquement des substantifs, ou des constructions de type N-de-N comme dans « *salle de bain* » sur l'exemple 3.5.

3.3.3 Intérêt des inférences pour la détection de polarité

Sur la version issue du consensus entre les quatre annotatrices, les 7972 portions annotées se répartissent en 6001 portions portant sur des mots ou des expressions, 1756 sur des phrases complètes, et 215 portions couvrant l'intégralité du commentaire. Nous présentons sur le tableau 3.1

deux types d'information en fonction de la valence (positive, négative) associée à la portion porteuse d'opinion : d'une part, si l'opinion exprimée est associée ou non à une inférence (présence / absence) ou s'il y a une incertitude à ce sujet, et d'autre part, pour les inférences jugées présentes, la répartition de ces inférences en termes de réalisation sémantique (logique, pragmatique, lexical) et mode de production (discursif, énonciatif).

Valence	Identification de l'inférence			Réalisation sémantique			Mode de production	
	absence	incertitude	présence	logique	pragma.	lexicale	discursif	énonciatif
Positif	1128	594	3222	1880	1476	2334	17	136
Négatif	575	1106	2996	2131	1875	1329	169	530
Total	1703	1700	6218	4011	3351	3663	186	686

TABLE 3.1 : Valence des opinions selon qu'elles sont associées à des inférences et répartition des inférences en fonction de la réalisation sémantique et des modes de production

Des opinions majoritairement exprimées dans un cadre inférentiel. En premier lieu, nous observons que toutes les opinions ne sont pas nécessairement associées à une inférence, mais elles sont cependant majoritairement exprimées dans le cadre d'une inférence, de manière sûre (64,6% des opinions dans le corpus, soit 6218 inférences). Le nombre d'opinions pour lesquelles il existe une incertitude sur la présence d'inférence pour les annotatrices humaines est aussi élevé que le nombre d'opinions exprimées en dehors d'une réalisation inférentielle. De plus, les opinions qui ne sont pas réalisées dans une inférence de manière certaine apparaissent nettement positives (66,2%, soit 1128 opinions).

Un usage équilibré entre les trois types de réalisation sémantique. En second lieu, nous constatons que lorsqu'une opinion est exprimée au moyen d'une inférence, la répartition entre les trois types de réalisation sémantique est globalement équilibrée, avec une légère majorité d'inférences logiques (36,4%), par rapport aux inférences lexicales (33,2%) et pragmatiques (30,4%). Nous émettons l'hypothèse que cette distribution s'explique par une recherche de compréhension par le plus grand nombre des avis postés sur internet, ce qui favorise les inférences logiques qui sont immédiates, au détriment des inférences pragmatiques qui reposent sur des connaissances externes telles que des connaissances culturelles. Enfin, bien que polarisées hors contexte, les inférences lexicales ne suffisent pas à exprimer seules une opinion, elles sont nombreuses mais elles apparaissent en complément ou dans une portion plus vaste qui aura été qualifiée de pragmatique (exemple 3.2), ce qui souligne l'importance de la prise en compte du contexte dans l'analyse des opinions. Néanmoins, ces inférences lexicales sont majoritairement associées à des opinions positives (63,7%) alors que les deux autres réalisations le sont avec des opinions négatives (à 53,1% pour les inférences logiques et 55,6% pour les inférences pragmatiques).

Des inférences très majoritairement réalisées dans un contexte énonciatif. En dernier lieu, nous constatons que les opinions exprimées au moyen d'inférences sont très majoritairement réalisées dans un contexte énonciatif (78,7%) plutôt que discursif. Cette observation renforce l'intérêt de tenir compte du contexte pour effectuer une fouille d'opinion lorsque l'opinion est réalisée dans une inférence.

3.3.4 Identification automatique de la polarité et des inférences

Ontologie des émotions

Nous avons appliqué sur le corpus de 1391 messages une ontologie des émotions en chinois [13] [Xu et al., 2008] fondée sur l'*Atlas des émotions* d'Ekman. Composée de 27 466 termes, elle associe à chaque terme une polarité et un score d'intensité. Un première observation montre qu'une minorité de commentaires (2099 avis, soit 29,3%) intègre des mots porteurs de sentiments. Nous avons utilisé l'information de polarité pour prédire automatiquement la valence globale de chacun des messages contenant au moins un mot porteur de sentiment. Parmi les commentaires dont la prédiction automatique de polarité est erronée, nous relevons qu'une majorité (79%) contient une inférence. Par ailleurs, s'il n'a pas été possible d'identifier un mot porteur de sentiment sur une majorité de messages (70,7%), la majorité de ces mêmes messages (81,9%) intègre au moins une inférence. Au final, une faible partie du corpus global (12,7%) ne se compose, ni de terme de sentiments, ni d'inférence, ce qui rend complexe l'analyse automatique de ces messages.

Nous pouvons tirer deux conclusions de cette observation. En premier lieu, le repérage d'inférences ne peut pas s'appuyer sur la seule présence de termes véhiculant des sentiments puisque la projection de l'ontologie ne couvre qu'une faible partie du corpus. En second lieu, il est raisonnable de tenir compte des inférences pour effectuer une fouille d'opinion puisqu'elle sont nombreuses en corpus et qu'elles véhiculent des opinions.

Classifieur automatique

Alors que l'annotation manuelle du corpus a été jugée complexe par les annotatrices, et que de nombreuses discussions ont été nécessaires pour aboutir à une version consensuelle, nous avons voulu vérifier les performances d'un classifieur automatique pour identifier les inférences. Liyun a entraîné six classifieurs SVM fondés sur des informations morpho-syntaxiques [14] et sur les méta-données [15] pour identifier la présence des inférences (premier modèle), puis, à partir de cette base, pour typer les inférences parmi cinq classes (un modèle par classe : logique, pragmatique, lexicale, énonciative, discursive) [Yan et al., 2020].

Sur le repérage des inférences (présence ou absence dans une chaîne de caractères), le premier modèle SVM obtient une précision élevée (0,919). Le tableau 3.2 présente les résultats obtenus par le classifieur sur l'identification des cinq types d'inférences retenus, autour des réalisations sémantiques et modes de production.

Evaluation	Identification des inférences	Réalisation sémantique			Mode de production	
		Logique	Pragmatique	Lexicale	Enonciative	Discursive
Précision	0,919	0,898	0,866	0,875	0,756	0,923

TABLE 3.2 : Performance du SVM en précision sur l'identification des inférences (présence ou absence) et sur le typage des inférences (réalisation sémantique et mode de production)

Nous observons que les résultats sont bons pour le repérage des trois types de réalisations sémantiques, avec des précisions supérieures à 0,866. Pour ce qui concerne le mode de production, la modalité énonciative obtient de moins bons résultats que la modalité discursive, alors que cette dernière est celle qui a été jugée la plus complexe par les annotatrices humaines.

13. http://ir.dlut.edu.cn/zyxz/qgbtk.htm ; l'ontologie intègre 5 375 termes neutres, 11 229 termes positifs, 10 783 termes négatifs, et 78 termes qui se polarisent en fonction du contexte.

14. Les informations morpho-syntaxiques utilisées sont de trois types : longueur du commentaire, nombre d'occurrences dans chaque classe de partie du discours par commentaire, et mots négatifs identifiés.

15. Les informations issues des méta-données sont de cinq types : score global de l'hôtel, nombre d'étoiles, score de confiance de l'utilisateur, âge de l'utilisateur, et localisation de l'hôtel.

Impact du système d'écriture sur la classification

Alors que le corpus après pré-traitements (voir page 18) combine des sinogrammes simplifiés et des séquences en pinyin pour les mots étrangers, nous avons souhaité vérifier si la conservation des sinogrammes était réellement utile pour l'identification des inférences. Pour cela, Liyun a converti l'ensemble du corpus en pinyin, sans les marqueurs de ton, en faisant l'hypothèse qu'il serait plus simple de traiter un corpus rédigé avec un nombre limité de signes distincts (les 26 caractères du système alphabétique latin) plutôt que de conserver le nombre important de sinogrammes (nous relevons 1675 sinogrammes simplifiés différents dans le corpus annoté de 1391 avis). Cette expérience a été réalisée en entraînant des modèles SVM globaux, permettant de typer les inférences parmi les cinq classes retenues (alors que l'expérience précédente s'appuyait sur des modèles unitaires par classe). Le tableau 3.3 permet une comparaison des performances du SVM sur le typage des inférences en cinq classes, selon que le modèle a été entraîné et appliqué sur la version du corpus avec sinogrammes, ou celle intégralement convertie en pinyin. Nous observons une très lé-

Evaluation	Réalisation sémantique			Mode de production		Macro
	Logique	Pragmatique	Lexicale	Enonciative	Discursive	Moyenne
Précision (sinogr.)	0,86	0,88	0,88	0,44	0,49	0,71
Précision (pinyin)	0,85	0,88	0,90	0,49	0,56	0,73

TABLE 3.3 : Comparaison des performances du SVM en précision sur le typage des inférences (réalisation sémantique et mode de production) sur les versions du corpus avec sinogrammes (haut) et intégralement convertie en pinyin (bas)

gère amélioration globale des résultats, avec une macro-précision qui évolue de 0,71 (version avec sinogrammes) à 0,73 (version en pinyin), et des résultats plus marqués sur le mode de production que sur la réalisation sémantique. Contrairement à ce que nous imaginions, les différences de résultats ne sont pas aussi marquées en passant de 1701 caractères (les 1675 sinogrammes complétés des caractères de l'alphabet latin pour les portions déjà rédigées en pinyin) à 26 caractères. En conséquence, et parce que la conversion en pinyin s'accompagne d'une perte d'information sémantique, nous concluons cette expérience par l'intérêt de conserver les sinogrammes dans les corpus de messages postés sur les réseaux sociaux en chinois.

3.4 Conclusion

Dans ce chapitre, j'ai présenté les recherches que j'ai menées avec Liyun Yan autour des inférences, dans un objectif d'amélioration des méthodes traditionnelles de fouille d'opinion. Ce travail a été réalisé sur un corpus d'avis touristiques de chinois en visite à Paris, rédigés en mandarin standard.

Alors que les recherches en linguistique entreprises par Charles Peirce donnent toujours lieu à discussion, une proposition de classification des inférences a été réalisée autour de trois niveaux d'analyse : la réalisation sémantique (lexicale, logique, ou pragmatique), la modalité de réalisation (par déduction, induction, ou par rétroduction), et la modalité de production (discursive ou énonciative). Aidée de plusieurs stagiaires chinoises étudiantes de l'INaLCO, Liyun a annoté un corpus de 1391 messages relatifs à 499 hôtels en s'appuyant sur cette classification des inférences. L'une des premières contributions que nous avons mis en évidence concerne l'intérêt des inférences pour la détection de la polarité, et plus généralement pour la fouille d'opinion, dans la mesure où ces inférences apparaissent dans les phrases ou portions dépourvues de termes porteurs d'opinion, sentiment, ou émotion. Cette observation trouve sa justification dans le traitement du chinois par les conventions sociales qui régissent les cultures asiatiques, comme le rappelle Benveniste (utilisation de périphrases pour éviter une attaque directe). L'identification des inférences s'avère donc complémentaire de la projection d'une ontologie des émotions.

Ayant démontré l'intérêt des inférences pour la fouille d'opinion, nous nous sommes ensuite intéressés à l'identification automatique des inférences et à leur typage. Plusieurs configurations à base d'apprentissage statistique (SVM) ont été réalisées : soit l'utilisation d'un modèle d'identification de la présence d'inférence dans un texte, suivie de l'utilisation de modèles pour chaque type d'inférence étudié, soit l'utilisation d'un modèle global qui permet de classer les inférences parmi les cinq classes retenues dans les expériences. Si les écarts sont faibles entre ces deux configurations au niveau des réalisations sémantiques, la deuxième approche d'un modèle global échoue à identifier correctement le mode de production. Cet échec doit cependant être relativisé par la distribution inégale des annotations entre classes, avec une sous-représentation assez marquée des annotations du mode de production par rapport à la réalisation sémantique d'une part, et par le faible nombre d'inférences discursives par rapport aux inférences énonciatives d'autre part.

Enfin, comme nous l'avons vu au chapitre 2, le traitement automatique du chinois est plus complexe que le TAL des langues européennes, en raison de l'absence d'espace dans les séquences de sinogrammes, et dont la segmentation en tokens peut potentiellement sur-segmenter un texte, mais également du fait de la coexistence de plusieurs systèmes d'écriture dans le même corpus (sinogrammes simplifiés, sinogrammes traditionnels, pinyin). En raison du nombre élevé de sinogrammes distincts (1675 sinogrammes simplifiés dans le corpus annoté), j'ai souhaité vérifier si la conversion des sinogrammes en pinyin ne permettait pas une amélioration notable des résultats, du fait de la réduction importante du nombre de caractères distincts. Contrairement à ma supposition d'origine, l'utilisation de modèles SVM entraînés sur un corpus à base de sinogrammes obtient des performances similaires à celle de modèles SVM entraînés sur un corpus intégralement converti en pinyin. Parce que la conversion automatique en pinyin s'accompagne d'une perte d'information sémantique, une dernière contribution dans ce travail de détection des inférences aura porté sur l'intérêt de conserver les sinogrammes simplifiés par opposition au pinyin.

Chapitre

Conclusion de la première partie

Dans cette première partie, je me suis intéressé à l'impact que les locuteurs d'une langue ont sur cette langue, en considérant la langue comme objet d'étude. J'ai orienté ma recherche sur les productions langagières réalisées principalement sur les réseaux sociaux, qui constituent une source d'informations régulièrement renouvelée et utile pour la fouille d'opinion et la pharmacovigilance, et dans une moindre mesure sur de la parole spontanée automatiquement transcrite.

J'ai étudié l'impact des locuteurs dans ces productions langagières en m'intéressant aux phénomènes présents dans la production numérique de locuteurs humains, tout en distinguant une hétérogénéité linguistique (vocabulaire, style, complexité sémantique) d'une hétérogénéité d'usage (émojis, hashtags, français inclusif). L'hétérogénéité linguistique peut s'imposer aux utilisateurs (translittération de mots étrangers) et dépend de l'origine géographique des locuteurs, comme dans le choix du système d'écriture (sinogrammes simplifiés en Chine continentale par opposition aux sinogrammes traditionnels encore utilisés à Hong Kong et Taïwan). Mais elle peut également relever d'un choix conscient des locuteurs, notamment par le biais de figures de style. L'hétérogénéité d'usage repose sur les fonctionnalités offertes aux utilisateurs sur les réseaux sociaux (hashtags,émojis), ou sur des choix politiques (productions langagières en français inclusif), voire une combinaison de ces deux aspects dans le cas des hashtags inclusifs.

Je me suis ensuite focalisé pendant la thèse de Liyun Yan sur les non-dits des locuteurs, dans un objectif d'amélioration de la fouille d'opinion. Ce travail a conduit à proposer une classification des inférences autour de trois niveaux d'analyse (réalisation sémantique, modalité de réalisation d'après les travaux de Peirce, et modalité de production), que nous avons utilisée pour annoter un corpus d'avis portant sur 499 hôtels (7972 portions annotées). J'ai pu observer que les opinions sont majoritairement exprimées dans un cadre inférentiel par les locuteurs de culture asiatique et qu'il est possible de les identifier automatiquement par des approches à base d'apprentissage. J'ai également constaté que l'identification de la polarité (positive/négative) est améliorée par ces informations. Enfin, j'ai pu vérifier que les systèmes automatiques obtiennent des performances similaires quel que soit le système d'écriture utilisé (sinogrammes simplifiés ou alphabet latin du pinyin).

Alors que cette première partie m'a permis de mettre en évidence la diversité des productions langagières des locuteurs d'une langue, et de leur impact sur la langue prise comme objet d'étude, j'aborde dans la deuxième partie de mon manuscrit l'impact que ces productions langagières ont sur les outils et ressources du traitement automatique des langues, en particulier au travers des modèles distributionnels.

Deuxième partie

Impact des utilisateurs sur les outils et ressources du Traitement Automatique des Langues

Chapitre 4

Du contenu utile en domaine médical

Sommaire

4.1 Introduction

Depuis mon arrivée au LIMSI en fin d'année 2006, une grande partie de mes activités de recherche a concerné le domaine médical, principalement dans le cadre de collaborations de recherche avec des CHU français, qui mettaient à notre disposition des textes cliniques produits par les équipes de soin (compte-rendu d'hospitalisation, certificats de décès). Sur la base de ces documents, j'ai principalement travaillé sur la thématique de la désidentification automatique [1] et sur l'extraction d'information fine comme les regroupements syndromiques. Ayant choisi d'orienter ce manuscrit d'HDR sur les liens entre le traitement automatique des langues et les productions langagières des locuteurs, j'ai écarté les travaux réalisés sur les documents produits par les médecins dans la mesure où ces contenus sont structurés [2] et normés (vocabulaire, formulations syntaxiques, indications d'intervalles numériques, etc.).

Mes activités dans le biomédical ont bénéficié, au niveau méthodologique, de mes participations aux campagnes d'évaluation internationales, en particulier la série des challenges i2b2 puis n2c2 autour de l'extraction d'information médicale en anglais [3] [Deléger et al., 2010, Minard et al., 2011, Grouin et al., 2011a, Zweigenbaum et al., 2012, Grouin et al., 2013a, Grouin, 2014a, Grouin, 2014b, Grouin et al., 2014b, Grouin et al., 2015, Grouin, 2016b, Grouin, 2016c, Zweigenbaum et al., 2016, Raithel et al., 2022], et dans une moindre mesure des campagnes BioNLP autour des mentions de bactéries et biotopes [Grouin, 2013b, Grouin, 2016a], CLEF eHealth sur la détection des maladies et leurs assertions [Hamon et al., 2014a], BioCreative sur l'identification des pathologies [Deléger et al., 2015], et

1. La désidentification consiste à repérer et traiter (masquer ou modifier) tout élément permettant d'identifier un patient dans une liste pré-définie de classes d'informations (nom, adresse, date, etc.). L'anonymisation va au-delà de la désidentification en ce qu'elle garantit que ce qui reste dans les textes cliniques ne permet pas de réidentifier le patient. Après mes travaux de thèse [Grouin, 2013a], j'ai orienté mes tr avaux vers la délexicalisation en générant des modèles sans information lexicale, contribuant à améliorer les performances puisque ces modèles sont moins dépendants des formes de surface, et permettant leur redistribution (https://github.com/grouin/medina/).

2. Les américains nomment la structure des compte-rendus hospitaliers au moyen de l'acronyme SOAP pour désigner les quatre grandes sections de base existantes : *Subjective, Objective, Assessment* et *Plan* [Podder et al., 2021].

3. Portail d'accès aux challenges n2c2 : https://n2c2.dbmi.hms.harvard.edu/challenges (éditions 2018 à 2022) et https://www.i2b2.org/NLP/RDoCforPsychiatry/PreviousChallenges.php pour les éditions 2008 à 2014.

SemEval sur les problèmes médicaux et les expressions temporelles [Grouin and Moriceau, 2016]. Dans la mesure du possible, j'ai tenté d'adapter au français les méthodes déployées sur l'anglais.

Au-delà de mes participations aux campagnes d'évaluation, j'ai participé à l'organisation du challenge CLEF eHealth sur la tâche de reconnaissance de concepts cliniques, puis sur le codage CIM-10 de ces concepts [Goeuriot et al., 2015, Névéol et al., 2015, Névéol et al., 2016, Névéol et al., 2017a, Névéol et al., 2017b]. Plus récemment, j'ai contribué à dynamiser ce champ de recherche pour le français, par le bais de l'organisation des campagnes d'évaluation DEFT [4], à partir de corpus de cas cliniques, et pour des tâches d'extraction d'information et de classification [Grabar et al., 2019, Cardon et al., 2020, Grouin et al., 2021]. Une différence majeure entre mes différentes contributions sur ces campagnes d'évaluation et les travaux que je présente dans la suite de cette section concerne le matériau pris en entrée : alors que les campagnes d'évaluation reposent sur des contenus produits par des professionnels du soin (cas cliniques, certificats de décès, compte-rendus hospitaliers, etc.), les travaux pour la pharmacovigilance ont porté sur les productions numériques de locuteurs sur les réseaux sociaux (voir chapitre 2).

Dans ce chapitre, j'ai donc décidé d'axer la présentation de mes travaux presque exclusivement sur la thématique de la pharmacovigilance, dans la mesure où ce domaine d'application m'occupe depuis 2014 avec plusieurs collaborations, d'abord en France auprès de l'Agence Nationale de Sécurité du Médicament et des produits de santé (projets Vigi4MED et Phares), puis à l'international (avec l'Allemagne et le Japon dans le cadre du projet tripartite Keepha).

4.2 Détection des effets secondaires

4.2.1 Les réseaux sociaux : une opportunité pour la pharmacovigilance

Développement et suivi des médicaments

Le processus de développement d'un médicament se compose de plusieurs phases, et commence par des essais en laboratoire pour déterminer la balance bénéfice-risque de la molécule testée, vérifier les effets secondaires, et déterminer la posologie adéquate. Une fois ces essais terminés et s'ils sont concluants, le médicament reçoit son autorisation de mise sur le marché (AMM). Le processus de pharmacovigilance consiste à poursuivre la surveillance d'un médicament et à prévenir tout risque d'effets secondaires, après qu'il a reçu son AMM et pendant toute la durée de vie du médicament [5]. Cette surveillance s'applique aux effets inattendus liés aux propriétés pharmacologiques à dose normale, ce qui exclu le mésusage (prise d'un médicament pour un autre usage ou sans respecter la posologie). Bien que des essais existent en laboratoire, il est impossible d'étudier tous les effets possibles résultant d'une interaction médicamenteuse, en raison du nombre élevé de produits dans la pharmacopée, ou lorsque les spécifications du médicament changent après son AMM (par exemple avec la nouvelle formule du Lévothyrox en mars 2017 [6]).

En France, la pharmacovigilance est assurée par l'Agence Nationale de Sécurité du Médicament et des produits de santé (ANSM) et un réseau de 31 centres de pharmacovigilance (CRPV) répartis sur l'ensemble du territoire, chargés de centraliser les déclarations d'effets secondaires. Les dispositifs mis en œuvre par les CRPV et les laboratoires pharmaceutiques pour permettre aux praticiens et aux patients de rapporter les effets indésirables se composent principalement de centres d'appel

4. DEFT (DÉfi Fouille de Texte) : https://deft.lisn.upsaclay.fr/2019/ (extraction d'information démographique : âge, genre, origine et issue de la consultation ou de l'hospitalisation ; similarité sémantique entre cas clinique et discussion, indexation des cas cliniques), https://deft.lisn.upsaclay.fr/2020/ (extraction d'information fine en dix classes autour de quatre domaines : patients, pratique clinique, traitements, temps) et https://deft.lisn.upsaclay.fr/2021/ (identification du profil clinique du patient).

5. https://solidarites-sante.gouv.fr/soins-et-maladies/medicaments/la-surveillance-des-medicaments/article/la-pharmacovigilance

6. https://ansm.sante.fr/actualites/levothyrox-levothyroxine-changement-de-formule-et-de-couleur-des-boites

et de formulaires en ligne [7]. Un rapport produit en 2014 par l'académie nationale de Pharmacie [8] révèle que seuls 4 à 5% des effets indésirables sont signalés de façon spontanée par les patients, et que la principale cause d'absence de déclaration est la méconnaissance des dispositifs de signalement. L'ANSM a constaté la même proportion de déclaration faite directement par les patients : en 2013, sur 46 843 effets secondaires déclarés, seuls 2 151 provenaient de patients, soit 4,6% ; en 2014, 1 983 déclarations de patients sur un total de 46 497, soit 4,3% seulement. C'est sur la base de ce constat que l'ANSM a décidé d'étendre ses activités de pharmacovigilance aux réseaux sociaux, et qu'elle a financé deux projets de recherche (Vigi4MED puis Phares) sur lesquels j'ai travaillé en lien avec les autres partenaires [Bœuf et al., 2017, Karapetiantz et al., 2018, Audeh et al., 2019].

Retours d'expériences

Cette approche semble confortée par la comparaison des informations identifiée sur le forum meamedica.fr spécialisé dans les retours d'expérience sur les médicaments (figure 4.1), par rapport à celles communiquées directement auprès des pharmacovigilants (figure 4.2).

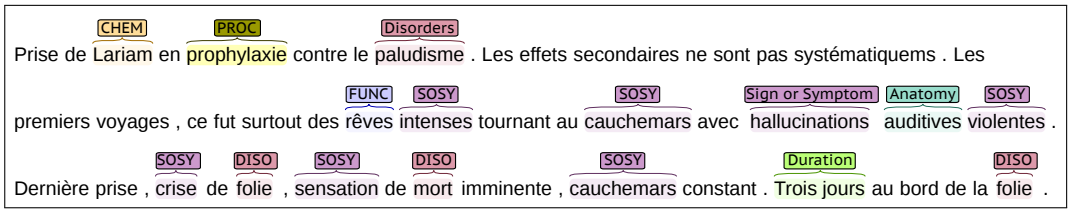

FIGURE 4.1 : Message d'utilisateur sur un forum de santé Meamedica.fr concernant les effets secondaires ressentis après la prise du Lariam

MEDICAMENT(S)

LARIAM

Lot	Voie	Dose	Fréquence	Du	Au	Durée	Délai surv	Dech	Rech	C	S	B	I	OMS
	ORL	1 DF						arrêt - 1	3	C2	S1	B3	I1	Suspect

Indication(s)

Prophylaxie du paludisme

EXAMENS COMPLÉMENTAIRES

Examen	Date	Valeur	Unité	Valeur normale	Classification

COMMENTAIRES

Description du cas:

Jeune femme de ans (DDN non précisée) qui déclare des effets indésirables survenus il y a lors d'un traitement préventif contre le paludisme par Lariam pour un voyage . Elle a arrêté précocément son traitement à 4 semaines en raison de la survenue d'effets indésirables à type de nausées, un épisode d'hallucinations, cauchemars, insomnie, crise d'angoisse.

Commentaires du notificateur :

Antécédent du patient :

Résultats d'examens complémentaires non structurés :

FIGURE 4.2 : Extrait anonymisé de compte-rendu de pharmacovigilance listant les effets secondaires ressentis par une personne ayant pris du Lariam

7. https://ansm.sante.fr/documents/reference/declarer-un-effet-indesirable
8. https://www.acadpharm.org/dos_public/GTNotif_Patients_Rap_VF___2015.01.22.pdf

Dans cet exemple relatif au Lariam, un traitement anti-paludéen pris de manière préventive (prophylaxie) avant un voyage, on retrouve dans les deux témoignages des notions de cauchemars, d'hallucinations, et de crises d'angoisse, qualifiées de *crises de folie* dans le message sur le forum de santé. Alors que le rapport officiel est normé (les termes de pathologies, signes et symptômes sont choisis), les messages sur les réseaux sociaux proposent des variantes et des précisions, notamment en termes de durée et fréquence (« *cauchemars constants* », « *trois jours au bord de la folie* »), et des indications assez personnelles qu'il n'est pas possible d'investiguer plus en détail faute de pouvoir revenir vers la personne concernée (« *sensation de mort imminente* »).

Si les informations d'ordre médical semblent pertinentes, il existe néanmoins une limite liée à l'impossibilité de revenir vers le patient pour obtenir des précisions (notamment sur les informations de posologie et d'interaction médicamenteuse ou sur la fréquence et l'intensité des phénomènes rapportés) ou se faire expliquer des expressions familières ou imagées (« *ralentissement du cerveau* », « *voir des schtroumphs* », « *avoir le QI d'une carotte* »). Au niveau TAL, traduire ces expressions en concepts cliniques se révèle complexe [9]. Malgré l'impossibilité de demander des informations complémentaires, l'identification du témoignage doit être considérée comme une détection du signal, qui permet au pharmacovigilant de prendre connaissance d'un fait, et doit l'inciter à interroger les autres CRPV pour vérifier l'existence de ce fait pour le valider.

Les corpus rassemblés pour les besoins de ces projets proviennent de 19 forums de santé en français [10], et dont les conditions générales d'utilisation autorisent la collecte des messages. J'observe que certains des forums de santé sur lesquels nous avons travaillé ont depuis fermé [11].

4.2.2 Impact des choix de schémas d'annotation

Comme indiqué au chapitre 2, la production des utilisateur est féconde mais nécessite des traitements dédiés en raison de la diversité des usages observée. Je me suis intéressé à l'identification des traitements et des effets secondaires, ainsi qu'aux relations existant entre les différentes entités observées en corpus. Une première partie de ce travail repose sur la définition des classes d'entités à identifier, autrement dit, à produire un guide d'annotation. J'ai exploré ce travail pendant le stage de Dalia Megahed [Megahed, 2014], puis du post-doctorat de François Morlane-Hondère.

Schéma d'annotation global

Dans une approche exploratoire [Grouin et al., 2014a], j'ai considéré quatre classes d'entités de haut niveau : les traitements médicamenteux, la posologie (dose, fréquence, durée), l'indication (ce pour quoi le traitement a été pris) et l'événement (effet secondaire positif ou négatif produit de manière inattendue par le traitement). Bien que les classes *indication* et *événement* renvoient à la même famille de concepts du point de vue médical (signe, symptôme, ou pathologie) voire aux mêmes formes de surface (« *je prends du XXX pour mon mal de tête* », indication, et « *chaque fois que je prends du XXX, j'ai un mal de tête* », effet secondaire), ma motivation pour proposer cette distinction repose sur les formulations libres employées par les utilisateurs pour décrire leur condition. Nous observons sur la figure 4.3 l'utilisation de termes et expressions qu'il est facile de normaliser dans une terminologie médicale (« *dépendant* ») et d'expressions longues qui nécessitent une interprétation (« *ton corps s'habitut* », « *ne te fais plus rien* ») et dont les termes qui composent ces expressions permettent difficilement d'identifier l'effet secondaire s'ils sont traités isolément.

9. Si « *voir des schtroumphs* » peut s'interpréter comme une hallucination visuelle, le « *ralentissement du cerveau* » renvoie t-il à l'absence de motivation ou à une fatigue intense ? Et « *avoir le QI d'une carotte* » correspond t-il à un manque de concentration ou à la difficulté à trouver ses mots ?

10. Ce corpus rassemble des forums généralistes sur les médicaments (Doctissimo), et des forums dédiés à une condition (grossesse), une maladie (diabète), ou un traitement particulier (baclofène).

11. Le médecin généraliste Dominique Dupagne justifie la fermeture du forum de son site atoute.org le 28 février 2022 par la perte de vitesse par rapport aux micro-blogs, le temps de gestion du forum (notamment les demandes parfois violentes de suppression de messages anciens), l'arrêt de l'indexation par les moteurs de recherche, etc (http://www.atoute.org/n/article405.html ?f=44). Concernant le forum baclofen.fr, c'est tout le site qui a fermé.

Ainsi, « *ton corps s'habitut* » et « *ne te fais plus rien* » renvoient à la notion d'accoutumance et correspondent indirectement au descripteur « Troubles liés à une substance » [12] dans le MeSH [13] tandis que « *dépendant* » correspond directement à ce même descripteur.

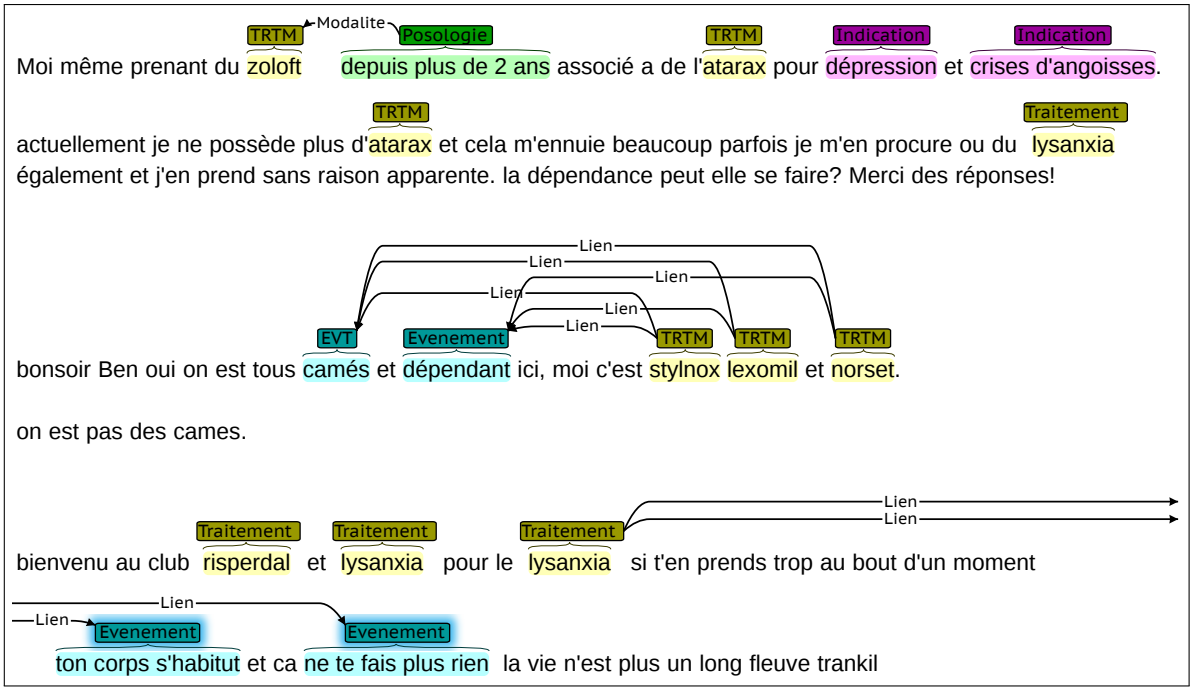

FIGURE 4.3 : Annotation de messages sur des forums de santé en quatre classes (Traitement, Posologie, Indication, Événement) avec relations typées entre certaines classes

Un corpus de quarante messages issus du forum sur la dépression a été annoté en double (totalisant 425 annotations de traitements, 221 posologies, 226 indications et 319 événements) pour vérifier la validité de ce schéma d'annotation. Les expériences d'apprentissage statistique à base de CRF de chaîne linéaire implémentés dans l'outil Wapiti [Lavergne et al., 2010] n'ont pas permis de valider ce schéma d'annotation. Sur un modèle mono-classe (prédiction des seuls effets secondaires) comme sur le modèle prédisant toutes les classes (modèle à quatre classes), le rappel des *événements* est très faible (voir tableau 4.1) et n'évolue que trop modérément en cas de fusion des classes *événement* et *indication* (modèle à trois classes).

Modèle	Mono classe			Quatre classes			Trois classes		
	R	P	F	R	P	F	R	P	F
Traitement	—	—	—	0,53	1,00	0,70	0,33	1,00	0,49
Posologie	—	—	—	0,22	0,65	0,33	0,07	0,67	0,13
Indication	—	—	—	0,23	0,64	0,33	0,14	0,67	0,23
Evénement	0,07	0,57	0,13	0,08	0,46	0,14			

TABLE 4.1 : Valeurs de rappel, précision et F-mesure sur la prédiction d'entités selon que le modèle a été entraîné sur une seule classe (événement), quatre classes (traitement, posologie, indication, événement) ou trois classes (après fusion indication/événement)

12. http://mesh.inserm.fr/FrenchMesh/view/loadSheet.jsp ?sheetId=D019966
13. Le MeSH (Medical Subject Heading) est un thésaurus rassemblant un vocabulaire contrôlé du domaine médical produit aux États-Unis par la National Library of Medicine (https://www.ncbi.nlm.nih.gov/mesh/), utilisé pour indexer les articles scientifiques dans PubMed. L'INSERM est chargée de la version française.

Les raisons expliquant ces mauvais résultats sur la classe la plus utile pour la pharmacovigilance sont multiples : la diversité des représentations du contenu de la classe *événement* ne permet pas de capturer aisément les propriétés du point de vue apprentissage, la faible taille du corpus appellerait de nouvelles annotations (longues et coûteuses) dans l'espoir d'améliorer les performances, et l'absence d'informations externes d'ordre médical utilisées pour l'apprentissage. Enfin, la thématique de la dépression semble plus complexe à traiter. Il a été observé que le recouvrement des entités de type *événements* et *indications* est plus important dans un corpus de témoignages sur les anti-dépresseurs que sur la migraine [Sarker et al., 2015]. En conséquence, j'ai abandonné ce schéma d'annotation global pour un schéma avec un niveau d'information plus fin.

Schéma d'annotation nucléaire

Dans le cadre du travail de post-doctorat de François Morlane-Hondère, nous sommes revenus à un schéma d'annotation à un niveau plus fin, avec seize classes d'entités autour de trois domaines :

— domaine du traitement : nom du traitement médicamenteux et informations de posologie (concentration, dose, mode)

— domaine clinique : signe ou symptôme, pathologie, partie anatomique, fonction biologique, procédure médicale, gènes/protéines

— domaine général : informations temporelles (date, heure, durée, fréquence) et complémentaires (poids, emploi)

Ces annotations sont qualifiées de nucléaires dans la mesure où l'élément atomique de ce schéma est le mot, et que les annotations produites ne portent généralement que sur un seul mot (voir figure 4.4). Nous avons appliqué ce schéma d'annotation nucléaire pour annoter un corpus de

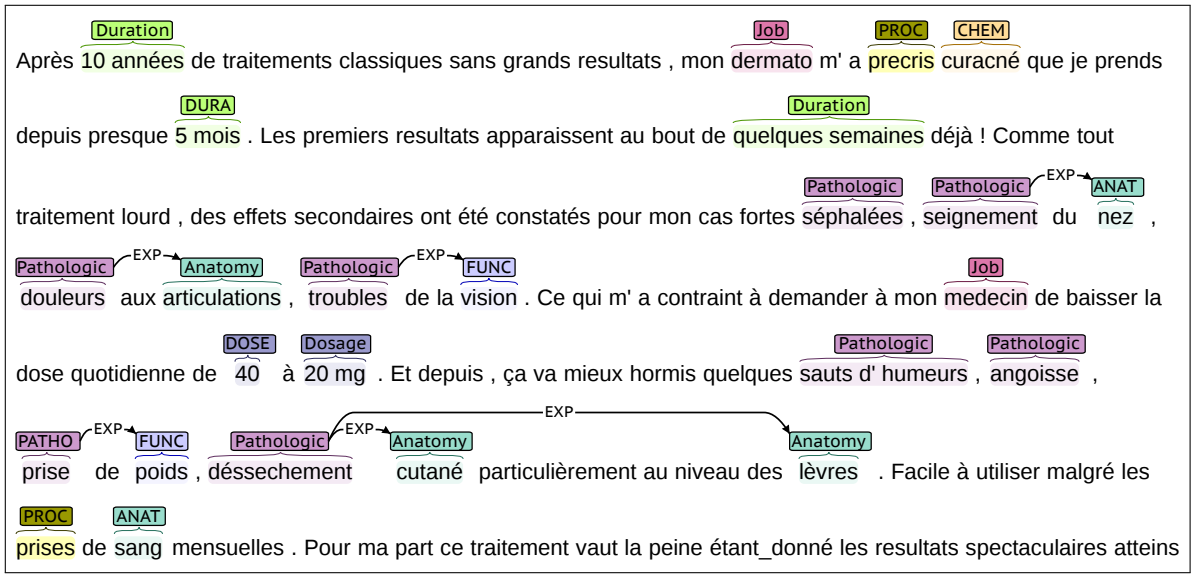

FIGURE 4.4 : Annotation de messages sur des forums de santé autour de 16 classes d'entités fines sur trois domaines (médicament, clinique, informations complémentaires) avec relations d'expansion entre certains types d'entités

1200 messages [Morlane-Hondère et al., 2016a, Morlane-Hondère et al., 2016b] extraits aléatoirement du site meamedica.fr. Cent fichiers ont été annotés en double avec un accord inter-annotateur qui valide le schéma (κ=0,825). Les 1100 messages suivants ont été annotés en simple annotation, avec une technique de *bootstrapping* pour simplifier le travail humain, par l'entraînement d'un modèle CRF sur les portions du corpus déjà annotées et corrigées, conduisant à une augmentation régulière des performances du modèle et une réduction de l'intervention humaine.

Comparaison Le tableau 4.2 présente les différentes classes d'annotation utilisées dans les deux schémas. La classe *Traitement* est commune aux deux schémas. Les trois informations de *Posologie* du schéma nucléaire sont rassemblées dans le schéma global. Concernant les informations temporelles, elles ne sont annotées dans le schéma global que si elles concernent la posologie et le traitement, alors qu'elles sont toujours annotées dans le schéma nucléaire en tant qu'informations complémentaires. Une différence majeure concerne les informations cliniques et anatomiques, réparties entre quatre classes dans le modèle nucléaire contre deux classes dans le schéma global, selon que ces informations concernent un *Evénement secondaire* ou une *Indication*. Enfin, six classes complémentaires ont été ajoutées au schéma nucléaire.

<table>
<tr><td>Schéma global</td><td>Traitement</td><td>Posologie</td><td colspan="2">Evénement | Indication</td><td>Absentes</td></tr>
<tr><td rowspan="2">Schéma nucléaire</td><td rowspan="2">Chemical</td><td>Concentration
Dosage
Mode</td><td colspan="2" rowspan="2">Sign or Symptom
Pathologic
Anatomy
Function</td><td>Procedure
Gene
Date
Time
Job
Weight</td></tr>
<tr><td>Duration
Frequency</td></tr>
</table>

TABLE 4.2 : Comparaison des classes d'entités utilisées dans le schéma d'annotation global à quatre classes (en haut : Traitement, Posologie, Evénement, Indication) et le schéma nucléaire à seize classes (en bas : Chemical, Concentration, Dosage, Mode, Duration, Frequency, etc.)

Une correspondance de schémas d'annotation est-elle possible ?

Sur cette base, j'ai souhaité vérifier s'il était possible de passer d'un schéma à un autre. J'ai effectué ce travail sur la problématique de la désidentification automatique, en m'appuyant sur les corpus en anglais de la campagne d'évaluation internationale i2b2/UTHealth NLP Challenge de 2014 qui portait sur la désidentification automatique [Stubbs et al., 2015], en ne considérant que les informations de personnes et de lieux. Bien qu'il ne s'agisse pas de contenus produits par les utilisateurs, les textes médicaux ont été produits par des professionnels du soin, j'ai préféré utiliser un schéma d'annotation défini par d'autres chercheurs que moi-même, en utilisant celui défini pour une campagne d'évaluation. Pour cela, j'ai comparé trois versions d'annotations :

— la version d'origine à sept classes *(médecin, patient, hôpital, adresse, ville, code postal, état)*, fondée sur des différences utiles pour l'humain mais complexes pour la machine (distinction médecin/patient)

— une version simplifiée en deux classes générales *(personne, lieu)* faisant consensus dans la communauté du REN

— et une version correspondant à une représentation plus régulière des entités à huit classes *(prénom, nom, code, numéro, ville, état, type, nom complémentaire)*, dans laquelle toutes les valeurs numériques sont rassemblées dans la même classe et où les distinctions de rôle (médecin/patient) sont écartées

Dans ce travail, mon objectif était de vérifier s'il était possible de transformer un schéma d'annotation fait par des humains pour le rendre utilisable par la machine, puis de revenir vers le schéma initial afin de fournir à l'humain ce qui était attendu [Grouin, 2018].

Les phrases suivantes donnent une représentation des correspondances entre schémas d'annotation sur les noms de personne (exemple 14) et les lieux (exemple 15), avec des boîtes encadrant les entités de la version d'origine en bleu (niveau intermédiaire), la version simplifiée encadrée en rouge (niveau extérieur), et la version régulière encadrée en vert (niveau intérieur).

(14) Ms. person[patient[first Michelle] last Klein] was seen in general neurology clinic today following her recent admission for complex migraine. Dr. person[doctor[first Remigio L.] last Allison] was present for all salient aspects of the history and physical exam.

(15) Internal Medicine.
loc[street[number 86] city Paris type Rd]
loc[city[city Washington]], loc[state[state DC]] loc[zip[number 20006]].

Expériences J'ai utilisé NeuroNER, fondé sur des réseaux de neurones récurrents (bi-LSTM), pour identifier les entités nommées, en conservant la configuration d'origine de l'outil [Dernoncourt et al., 2017] (tokénisation avec spaCy, plongements lexicaux fournis et entraînés avec GloVe, optimisation SGD). La figure 4.5 présente l'évolution de la F-mesure sur le corpus de test, pour chaque version du schéma. Le tableau 4.3 présente les résultats par catégorie à l'issue de la construction du modèle. Le tableau 4.4 présente les résultats de NeuroNER suivi de règles de conversion des types prédictions d'entités nommées (sans modification de résultats sur la v1 puisque les types d'entités prédits sont déjà ceux attendus). Le tableau 4.5 présente les résultats lorsque la conversion est appliquée sur les entités de référence. Cette évaluation met en évidence la qualité des règles de conversion indépendamment de l'identification des entités nommées.

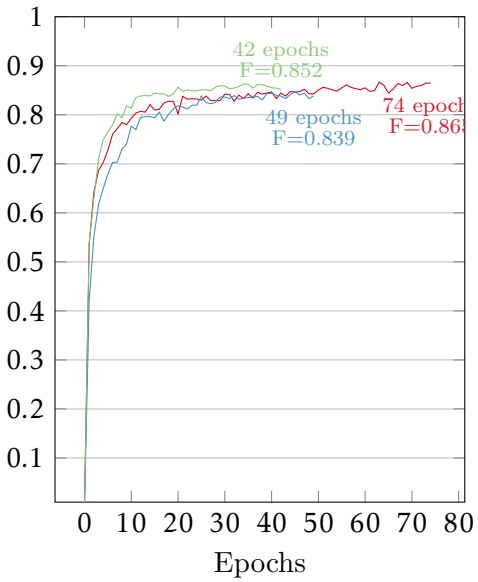

FIGURE 4.5 : Évolution de la F-mesure sur le test par itération pour les versions d'origine (v1), simplifiée (v2) ou régulière (v3)

Type	v1 (origine)			v2 (simple)			v3 (régulier)		
	R	P	F	R	P	F	R	P	F
Pers.	–	–	–	.885	.938	.911	–	–	–
Doc.	.859	.929	.893	–	–	–	–	–	–
Pat.	.846	.863	.854	–	–	–	–	–	–
First	–	–	–	–	–	–	.887	.919	.903
Last	–	–	–	–	–	–	.917	.927	.922
Code	–	–	–	–	–	–	.657	.888	.755
Loc.	–	–	–	.737	.836	.784	–	–	–
Hosp.	.655	.830	.732	–	–	–	–	–	–
Street	.838	.826	.832	–	–	–	–	–	–
City	.633	.788	.702	–	–	–	.633	.761	.691
State	.790	.802	.796	–	–	–	.756	.791	.773
Zip	.914	.928	.921	–	–	–	–	–	–
Num.	–	–	–	–	–	–	.834	.931	.883
Name	–	–	–	–	–	–	.488	.743	.589
Type	–	–	–	–	–	–	.842	.879	.860
TOUS	.800	.881	.839	.831	.902	.865	.818	.888	.852

TABLE 4.3 : Évaluation (rappel, précision, F-mesure) du repérage d'entités nommées selon le schéma d'annotation utilisé pour créer le modèle

La figure 4.5 met en évidence qu'un schéma d'annotation complexe (v1 avec 7 types) obtient de moins bons résultats que la version simplifiée (v2 avec 2 types) tandis qu'un schéma plus régulier (v3 avec 8 types) permet à NeuroNER d'obtenir plus rapidement de meilleurs résultats (la courbe verte augmente plus rapidement que les autres). La distinction entre patient et médecin

Cat.	v1 (origine)			v2 (simple)			v3 (régulier)		
	R	P	F	R	P	F	R	P	F
Doc.	.859	.929	**.893**	.848	.800	.823	.788	.673	.726
Patient	.846	.863	**.854**	.713	.896	.794	.473	.447	.460
Hosp.	.655	.830	**.732**	.591	.611	.601	.459	.634	.532
Street	.838	.826	**.832**	.596	.853	.701	.588	.370	.455
City	.633	.788	**.702**	.617	.593	.605	.481	.418	.447
State	.790	.802	.796	.795	.853	.823	.816	.856	**.836**
Zip	.914	.928	.921	.900	.984	**.940**	.900	.984	**.940**
TOUS	.800	.881	**.839**	.747	.774	.760	.640	.607	.623

TABLE 4.4 : Rappel, précision, et F-mesure des prédictions de NeuroNER avec post-traitement

v1 (origine)			v2 (simple)			v3 (régulier)		
R	P	F	R	P	F	R	P	F
.859	.929	**.893**	.950	.834	.888	.879	.702	.780
.846	.863	**.854**	.735	.896	.808	.507	.471	.488
.655	.830	.732	.830	.724	**.774**	.753	.747	.750
.838	.826	**.832**	.721	.961	.824	.794	.480	.598
.633	.788	**.702**	.761	.638	.694	.883	.581	.701
.790	.802	.796	.842	.856	.849	.953	.924	**.938**
.914	.928	.921	.943	.971	.957	.957	.971	**.964**
.800	.881	**.839**	.860	.815	.837	.664	.783	.718

TABLE 4.5 : Rappel, précision, et F-mesure de conversions sur les entités de référence

produit de moins bons résultats que celle entre prénom et nom (tableau 4.3), mettant en évidence cette complexité pour un système statistique. Malgré la réduction de l'ambiguïté dans le schéma d'annotation plus régulier, nous obtenons des résultats plus faibles pour les types d'EN utilisés dans plusieurs contextes (F=0.691 vs. 0.702 pour les villes, F=0.773 vs. 0.796 pour les États) par rapport au schéma d'origine. Puisque les systèmes statistiques sont sensibles aux variations contextuelles, une plus grande variété de contextes dans lesquelles apparaissent des EN de ces types a un impact négatif sur les performances du système. De manière non intentionnelle, nous avons remplacé l'ambiguïté de définition par une ambiguïté de contexte. La conversion des types d'EN depuis les deux versions simplifiées se révèle complexe dans la mesure où elle revient à réintroduire de l'ambigüité dans un jeu d'annotations simplifiées.

4.2.3 Relations d'expansion entre entités pour retrouver les concepts d'intérêt

Si le schéma nucléaire permet d'identifier facilement les portions contenant des entités en raison de leur taille réduite (un seul terme), il importe ensuite de relier entre elles les entités nucléaires. Notre approche repose sur des relations d'expansion entre entités [Morlane-Hondère et al., 2016a]. Nous avons manuellement annoté les relations sur le corpus Meamedica de 1200 messages (comme entre *douleurs* et *articulations* sur la figure 4.4) pour quatre classes d'entités *(Anatomie, Fonction, Pathologie, Signe ou Symptôme)* et dans la limite de sept mots entre deux entités. Nous avons ensuite entraîné un modèle avec l'outil LibSVM [Chang and Lin, 2011], en fournissant le nombre de mots entre les deux entités liées, les parties-du-discours des entités et des mots entre entités, et les types d'entités des deux entités. Le corpus d'entraînement se compose des relations annotées manuellement (exemples positifs) et de fausses relations générées aléatoirement (exemples négatifs). Nous avons généré toutes les relations d'expansion possibles sur le corpus de test, le modèle devant déterminer celles qui correspondent à de réelles relations d'expansion. Le tableau 4.6 présente les résultats obtenus par le modèle SVM sur le corpus de test (160 relations d'expansion à identifier), selon que le modèle est appliqué sur les entités de référence (bas du tableau)

Entités en entrée	Approche	P	R	F
Prédictions du CRF	SVM	0,683	0,956	0,797
	SVM + règles	0,746	0,881	0,808
Annotations de référence	SVM	0,724	0,969	0,829
	SVM + règles	0,770	0,900	0,830

TABLE 4.6 : Valeurs de rappel, précision, et F-mesure sur l'identification des relations d'expansion à partir des entités de référence (bas du tableau) ou d'entités identifiées par un modèle CRF (haut du tableau), avec et sans règles de post-traitement des relations prédites

ou préalablement identifiées par un modèle CRF avec l'outil Wapiti [Lavergne et al., 2010] (haut du tableau), avec et sans règles de post-traitement. Ces règles sont simples et empiriques ; par exemple, pas de relation d'expansion entre deux entités séparées par des marques de ponctuation forte (point, deux-points).

Nous constatons que les résultats sont assez proches dans l'identification des relations d'expansion, que cette identification ait été faite sur les entités de référence (F=0,829) ou sur celles prédites par le modèle CRF (F=0,797). Nous observons également que les règles de post-traitement n'ont permis qu'une faible amélioration, principalement marquée au niveau de la précision, mais que cette amélioration permet de rééquilibrer les différences entre vrais positifs et faux positifs. Pour autant, l'approche envisagée consistant à discriminer des relations d'expansion parmi toutes les relations potentielles générées entre entités, est validée par les bonnes performances obtenues.

4.3 Des indices linguistiques pour améliorer la détection des informations

Afin d'améliorer l'identification des informations pertinentes pour la pharmacovigilance, nous nous sommes intéressés aux indices linguistiques qu'il est possible d'utiliser, soit comme approche de post-traitement après avoir identifié les traitements médicamenteux, signes, symptômes, pathologies, soit pour fournir des caractéristiques d'apprentissage supplémentaires. Nous avons travaillé sur le typage des relations causales entre médicament et pathologie au moyen de connecteurs logiques, sur les verbes introducteurs de médicaments, sur le repérage des variantes et erreurs orthographiques, et enfin, sur l'utilisation de voisins distributionnels pour identifier des candidats pour normaliser un concept cible. Plus spécifiquement en raison d'une actualité forte en pharmacovigilance, j'ai également travaillé à la détection du mésusage sur des cas d'usage précis, puis lors de la pandémie de Covid-19. Pour cette dernière, je n'ai aucun résultat à présenter dans ce manuscrit, les données traitées n'ayant pas permis d'identifier des informations utiles.

4.3.1 Typage des relations causales

Dans le but de distinguer les *indications* des *événements secondaires*, j'ai travaillé avec François sur le typage des relations causales [Morlane-Hondère et al., 2015a]. Nous nous sommes appuyés sur les connecteurs logiques provenant du lexique de 231 connecteurs Lexconn [Roze et al., 2012, Roze, 2013], en excluant les connecteurs ambigus. En raison de la complexité induite par la thématique de la dépression, nous nous sommes focalisés sur le Médiator [14] dont les effets secondaires sont connus et documentés (*troubles digestifs, fatigue, vertiges,* etc.). L'exemple 16 fournit un exemple d'annotations réalisées avec le schéma nucléaire, en distinguant les entités relevant de l'*indication* en bleu, de celles associées à un *événement* en rouge, tandis que celles en vert ne dépendent d'aucun de ces deux rôles. Le type d'entité est indiqué en exposant.

(16) Suite à quelques sosy:malaises avec sosy:perte de func:connaissance, mon job:endocrinologue m'a proc:prescris du chem:Médiator. Au début c'est vrai j'ai eu le phénomène sosy:perte de func:poids (weight:6kg). Au fil des années, diso:migraines, diso:diarhées, diso:crampes, une sosy:fatigue de plus en plus grande, une sosy:hyper-émotivité, toujours sur les anat:nerfs.

Nous avons observé qu'un tiers des *indications* sont introduites par un marqueur de but (« *pour, afin de, dans le but de* ») qui exprime la relation entre la prise d'un médicament et le but visé par cette prise. Les marqueurs d'explication (« *car, dans le cadre, pour cause de* ») jouent un rôle similaire, mais leur différence d'emploi entre *indication* et *événement* est moins marquée. Enfin, nous avons constaté que la relation d'opposition (« *or* ») est potentiellement intéressante

14. Hypoglycémiant utilisé par les diabétiques pour lutter contre une glycémie excessive, et détourné de son usage par les non-diabétiques pour perdre du poids.

pour introduire un phénomène inattendu (exemple 17). De cette étude, il ressort que ces indices permettent plus facilement de repérer les *indications* que les *événements*.

(17) mon endocrinologue me prescrit du LEVOTHYROX 75mg par jour + 3MEDIATOR. or je prends de plus en plus de poids

4.3.2 Verbes introducteurs de médicaments

Nous nous sommes ensuite intéressés aux verbes introducteurs de médicaments (VIM) [Morlane-Hondère et al., 2015b], toujours dans l'objectif de distinguer ce qui justifie la prise d'un médicament de ce qui est causé par cette prise. Dans une première étape, nous avons constitué un lexique de verbes introducteurs de médicaments en nous fondant sur la présence du pronom personnel « *je* » et d'un nom de médicament, dans une fenêtre de six mots. La pertinence des verbes ainsi identifiés a ensuite été évaluée manuellement. Dans une deuxième étape, nous avons utilisé cette liste de VIM pour identifier de potentiels noms de médicaments complémentaires au moyen d'une règle basique : un pronom personnel suivi d'un VIM suivi d'un mot absent d'une liste des formes fléchies du français, dans la limite de six mots après le verbe. L'application de cette règle nous a permis de mettre en évidence des erreurs orthographiques en termes d'insertion de caractère (« **cortisonne* » au lieu de « *cortisone* »), de délétion (« **lévotyrox* » vs. « *lévothyrox* »), de substitution (« **allupirinol* » vs. « *allopirinol* »), d'inversion (« **steridil* » vs. « *stediril* »), ou d'une combinaison de ces éléments (« **calements* » vs. « *calmants* »).

Une première conclusion de ce travail concerne la sémantique des verbes. Nous distinguons quatre groupes de verbes en fonction du type d'information véhiculé : information temporelle *(commencer, continuer, essayer, tenter)*, information de dosage *(augmenter, baisser, diminuer)*, information sur la quantité excessive *(abuser, bouffer, manger)*, ou information galénique *(avaler, se badigeonner, boire, sucer, sniffer)*, avec certains emplois familiers liés à l'expression sur les réseaux sociaux. Une deuxième conclusion plus inattendue concerne le lien entre certains verbes et le type de traitement introduit : « *prescrire* » pour les anti-dépresseurs *(Atarax, Lexomil, Lysanxia, Xanax)*, « *prendre* » et « *donner* » pour les anti-douleurs sans ordonnance *(Doliprane, Gaviscon, Spasfon)*, « *avaler* » pour les médicaments connus et rassurants *(Doliprane)*, et « *commencer* » pour les traitements de longue durée contre la dépendance aux opiacés *(Subutex)*.

4.3.3 Détection de variantes et de fautes d'orthographe

Pour faire suite aux verbes qui introduisent des noms de médicaments, même si le nom de médicament est mal orthographié, nous avons poursuivi ce travail par la recherche de variantes et l'identification d'erreurs orthographiques [Morlane-Hondère et al., 2016a], dans l'objectif de fournir des caractéristiques d'apprentissage supplémentaires lors de la constitution de modèles statistiques. Sur la base du corpus de 1200 messages du site meamedica.fr annotés selon le schéma d'annotation nucléaire (p. 54), nous avons appliqué l'outil word2vec [Mikolov et al., 2013] et généré quatre sorties en fonction du nombre de clusters demandé (100, 200, 400 et 800). Si les identifiants de cluster ainsi produits peuvent servir de caractéristique d'apprentissage, le résultat permet également de mettre en évidence des variantes et des erreurs orthographiques. Par cette approche, nous avons pu identifier jusqu'à 19 variantes pour le terme « *gynécologue* », telles que la version raccourcie « *gynéco* », la version familière « *gygy* », ou encore des fautes comme « *génico* ».

4.3.4 Voisins distributionnels pour la normalisation

Au-delà des variantes, nous avons également travaillé à la normalisation de concept en nous appuyant sur les plongements de mots générés par word2vec et la recherche de voisins distributionnels. Cette approche est utilisée pour la désambiguïsation sémantique [Gallant, 1991, Navigli, 2009], pour construire des ressources linguistiques [Claveau et al., 2014, Ferret, 2015]. En domaine biomédical, elle a été appliquée pour réduire la diversité des termes utilisés en corpus [Périnet and Hamon, 2014] ou pour de l'expansion d'abréviations [Wu et al., 2015], mais

également sur une tâche de normalisation de concepts cliniques dans l'UMLS lors de la campagne SemEval 2014 [Kaewphan et al., 2014]. L'originalité de l'approche que nous avons suivie [Morlane-Hondère and Grouin, 2016] repose sur le parcours des représentations vectorielles générées, en faisant varier la taille du contexte (fenêtres de taille 1, 5, 10 et 20), parmi trois parcours :

— le parcours en profondeur consiste à parcourir les vingt premiers voisins d'une entité à normaliser, et à afficher ces voisins ; cette approche permet d'identifier les meilleurs voisins par score de similarité décroissant

— le parcours en largeur consiste à parcourir, pour chacun des quatre premiers voisins, les quatre premiers voisins qui lui sont associés (les voisins des voisins) et à les afficher, puis à réitérer pour chaque voisin de premier niveau ; cette approche repose sur l'idée de maximiser les meilleurs voisins en accédant aux voisins du deuxième degré, plutôt qu'en parcourant l'intégralité des voisins de premier degré

— le parcours mixte consiste à d'abord parcourir les quatre premiers voisins, avant de parcourir les quatre premiers voisins de chacun des voisins de premier niveau

Nous reportons dans le tableau 4.7 la liste des voisins parcourus pour le mot cible « *pertes d'audition* » au moyen du modèle appris sur une fenêtre de taille 1. Les voisins identifiés dans l'UMLS [Lindberg et al., 1993] sont notés avec la mention « (u) », et correspondent à des normalisations potentielles. Nous avons souligné les voisins que nous avons considérés comme relevant de normalisations valides.

parcours en profondeur	parcours en largeur	parcours mixte
PERTES D'AUDITION hyperacousie (u) acouphéniques acouphènes (u) pertes auditives (u) trauma sonore acouphénique otospongiose (u) perte d'audition baisse d'audition perte auditive (u) maladie de ménière (u) acouphénes sifflements (u) traumatisme sonore (u) acouphènique hyperacousique surdité brusque (u) bourdonnements acouphenes accouphènes	PERTES D'AUDITION hyperacousie (u) acouphènes (u) acouphénique acouphéniques trauma sonore acouphéniques hyperacousie (u) acouphénique acouphènes (u) acouphénes acouphènes (u) hyperacousie (u) acouphéniques sifflements (u) acouphène (u) pertes auditives (u) hyperacousie (u) otospongiose (u) perte d'audition acouphénique	PERTES D'AUDITION hyperacousie (u) acouphéniques acouphènes (u) pertes auditives (u) hyperacousie (u) acouphènes (u) acouphénique acouphéniques trauma sonore acouphéniques hyperacousie (u) acouphénique acouphènes (u) acouphénes acouphènes (u) hyperacousie (u) acouphéniques sifflements (u) acouphène (u) pertes auditives (u) hyperacousie (u) otospongiose (u) perte d'audition acouphénique

TABLE 4.7 : Voisins ramenés pour le mot cible « *pertes d'audition* », suivis de (u) pour ceux présents dans l'UMLS, soulignés pour ceux correspondant à des normalisations valides

Si le parcours en profondeur est le plus intuitif, le parcours en largeur, parce qu'il se limite aux voisins de premier rang, permet de retrouver le plus de candidats probables, mais aussi de ramener potentiellement plusieurs exemplaires du même candidat. En combinant ces deux parcours, les voisins examinés sont les mêmes que ceux du parcours en largeur, la différence résidant dans l'ordre du parcours. En raison des différences de parcours, nous observons que la normalisation « *pertes auditives* » que nous avons jugée valide est en quatrième place dans le modèle mixte alors qu'elle est seizième dans le modèle en largeur. Nous remarquons également que le parcours des voisins de deuxième degré entraîne la présence de doublons dans les voisins rapportés.

Une première conclusion de ce travail est qu'il n'est pas possible d'identifier un type de parcours meilleur qu'un autre pour la normalisation. Si le mode en profondeur permet d'identifier moins de candidats, ceux identifiés sont pertinents pour la normalisation. Une deuxième conclusion révèle que les modèles avec une taille restreinte (un seul mot) ou trop large (vingt mots) sont parmi les moins performants dans nos expériences.

4.4 Détection du mésusage médicamenteux

4.4.1 Présentation

Dernièrement, je me suis intéressé à la détection du mésusage médicamenteux rapporté sur les réseaux sociaux [Campillos-Llanos et al., 2019] à l'occasion du post-doctorat de Leonardo Campillos-Llanos. L'ANSM définit le mésusage comme un « *usage intentionnel et inapproprié* » d'un médicament ou d'un produit de santé, qui « *n'est pas conforme à l'autorisation de mise sur le marché (AMM) ou aux recommandations de bonnes pratiques* ». Au niveau TAL, il s'agit d'abord d'identifier et de vérifier l'intentionnalité d'un locuteur dans ses messages, sachant que les effets secondaires augmentent de 50% en cas d'utilisation d'un médicament hors-AMM [15]. Si les enjeux de santé publique apparaissent nettement dans ce pourcentage, la nécessité d'obtenir rapidement un traitement médicamenteux contre la pandémie de Covid-19 a également montré son intérêt pour le mésusage rapporté sur les réseaux sociaux. Tout traitement et bénéfice supposé contre les symptômes rapportés sur les réseaux sociaux constituaient des perspectives d'investigation.

Les pharmaciens distinguent plusieurs types de mésusage : non-respect de la posologie (dose, durée, fréquence) ou des recommandations de prise (à jeûn, pendant le repas), utilisation d'un médicament pour une autre pathologie que celle prévue dans l'AMM ou pour un usage récréatif (cocktail *Purple drank* à base de sirop contre la toux contenant de la codéine). La détection du mésusage est encore plus complexe que celle des effets indésirables dans la mesure où il s'agit d'identifier des traitements médicamenteux qui ont été pris intentionnellement dans un objectif qui ne correspond pas aux usages définis dans l'AMM. En raison de l'absence de données annotées disponibles sur le sujet, nous avons opté pour des approches non-supervisées. Par opposition aux méthodes de supervision distantes qui ont pu être employées dans le médical [Mintz et al., 2009, Poon et al., 2014], nous avons privilégié l'approche consistant à annoter des sources de connaissances proches du domaine pour ensuite les appliquer sur les forums de santé.

4.4.2 Preuve de concept

Corpus et annotations Les pharmacovigilants de l'HEGP ont annoté la présence de mésusage dans un corpus de 1178 messages relatifs à cinq traitements médicamenteux dont le mésusage est connu [16]. Des formes de mésusage ont été identifiées sur 111 messages, confirmant l'utilité des témoignages d'utilisateurs pour la veille sanitaire, avec quelques formulations révélatrices :

15. https://ansm.sante.fr/page/lidentification-et-le-traitement-des-signaux

16. L'Agomélatine et la Duloxétine sont des anti-dépresseurs détournés pour combattre les insomnies ou les crises de panique et d'anxiété. Le Baclofène est un relaxant musculaire utilisé contre les addictions à l'alcool. Le Myolastan est également un relaxant musculaire connu pour le mésusage en termes de posologie et d'indication. Enfin, l'Exénatide est un anti-diabétique dont le sur-dosage a été constaté par les pharmacovigilants.

— Mésusage : « *surdosé* », « *mauvais dosage* », « *tu ingurgites le triple de ce qui est recommandé* », « *prendre des doses de cheval de XXX* »

— Indication : « *je prends XXX pour* », « *XXX utilisé pour* », « *XXX utilisé comme* »

— Effet secondaire : « *XXX m'empêchait de dormir* »

Pour évaluer la pertinence de la méthode, nous avons rassemblé dans le corpus de test des messages relatifs à trois cas d'usage en rapport avec l'actualité, à raison de cent messages par cas : le Baclofène (permettant une évaluation directe par rapport aux données d'entraînement), la Lévothyroxine (en raison du changement de formule introduit en août 2017), et la vaccination (la France est passée au 1er janvier 2018 de trois à onze vaccins obligatoires [17] chez les nourrissons). Pour constituer la référence du mésusage, nous avons utilisé les résumés des caractéristiques du produit (RCP) disponibles depuis la base de données publique du médicament [18] pour les médicaments correspondants aux trois cas d'usage, en nous focalisant sur les sections relatives aux indications, contre-indications, et effets secondaires.

Méthode Reprenant la méthode élaborée par [Bigeard et al., 2018] sur la détection et la classification du mésusage en français sur les réseaux sociaux, nous avons entraîné un modèle bayésien naïf en validation croisée 10-plis. Les caractéristiques d'apprentissage reposent sur les tokens, la racine des mots, les codes ATC [19] et MedDRA [20] des pathologies (par une recherche à l'identique), et des trigrammes de tokens et de caractères. Pour vérifier l'impact du déséquilibre entre classes sur les performances, nous avons défini trois ratios de messages avec et sans mésusage (le ratio initial de 10:1, soit 1178 messages ; un ratio de 2:1 avec deux fois plus de messages sans mésusage, soit 346 messages ; et un ratio équilibré entre mésusage et sans mésusage totalisant 246 messages). Les entités du corpus des classes *Traitement* et *Pathologie* ont été annotées par un modèle CRF dont les performances en corpus (forums et RCP) sont données dans le tableau 4.8. Nous observons un équilibre du nombre d'entités de référence entre classes sur chaque corpus (1089 traitements et 1160 pathologies sur les forums, contre 3085 traitements et 3687 pathologies dans les RCP).

Corpus	Pathologies			Traitements			R	P	F
	R	P	F	R	P	F			
Forums	0,70	0,83	0,76	0,83	0,93	0,88	0,77	0,88	0,82
RCP	0,90	0,92	0,91	0,90	0,76	0,82	0,90	0,84	0,87

TABLE 4.8 : Performance du modèle CRF sur la détection des entités de type Pathologies et Traitement sur les forums de santé et les résumés de caractéristiques du produit (RCP)

Le tableau 4.9 présente les résultats d'identification du mésusage rapporté sur des forums de santé, au moyen d'un classifieur bayésien naïf, en fonction des trois ratios précédemment définis.

Ratio	R	P	F
10:1	0,93	0,94	0,94
2:1	0,67	0,63	0,65
1:1	0,88	0,88	0,88

TABLE 4.9 : Performances du modèle bayésien naïf sur la détection du mésusage en forum de santé, selon le ratio de messages avec/sans mésusage retenu

17. Diphtérie, tétanos, poliomyélite, complétée par coqueluche, rougeole, rubéole, oreillons, hépatite B, et bactéries *haemophilius influenzae B*, *méningocoque C*, et *pneumocoque* : https://solidarites-sante.gouv.fr/prevention-en-sante/preserver-sa-sante/vaccination/vaccins-obligatoires/article/11-vaccins-obligatoires-depuis-2018
18. https://base-donnees-publique.medicaments.gouv.fr/
19. Classification ATC (Anatomical Therapeutic Chemical) : https://www.whocc.no/atc_ddd_index/
20. MedDRA (Medical Dictionary for Regulatory Activities) : https://www.meddra.org/

Une première contribution de cette étude concerne l'intérêt d'utiliser les résumés de caractéristiques du produit (RCP) en complément des contenus disponibles sur les réseaux sociaux, de manière à élargir le nombre de ressources utiles ou pour produire une référence en l'absence de données annotées. La qualité de ces données, au regard des valeurs élevées en terme de rappel sur l'identification des entités, confirme l'intérêt de ces données. Une deuxième contribution confirme l'intérêt de la méthode suivie par [Bigeard et al., 2018] pour identifier le mésusage sur les forums de santé. Nous constatons également que rééquilibrer le nombre de messages avec et sans mésusage pour l'entraînement ne semble pas pertinent, puisque nos meilleurs résultats ont été obtenus en conservant le ratio fortement déséquilibré d'origine. Enfin, ce travail exploratoire mérite d'être poursuivi, en combinaison avec les travaux effectués par François, en particulier en utilisant les verbes introducteurs de médicaments.

4.5 Conclusion

Dans ce chapitre, j'ai présenté l'état de mes recherches autour des activités de pharmacovigilance sur les réseaux sociaux, durant le stage de Dalia Megahed pour mettre en place un schéma d'annotation global (p. 52), puis du post-doctorat de François Morlane-Hondère pour retravailler ce schéma avec une version nucléaire où les éléments atomiques sont désormais des mots simples (p. 54) reliés par des relations d'expansion qu'il est possible d'obtenir par un SVM (p. 57). Les expériences à base de plongements de mots et de voisins distributionnels nous ont permis de progresser dans plusieurs directions, notamment pour identifier des variantes de concepts (p. 59) ou pour permettre une normalisation au travers de plusieurs parcours possibles des termes candidats (p. 59). Dernièrement, je me suis intéressé avec Leonardo Campillos-Llanos au mésusage rapporté sur les réseaux sociaux, que nous avons exploré au moyen de classifieur bayésien naïf (p. 61).

Au-delà des formulations propres à chaque locuteur dans les productions langagières sur les forums de santé, et qui supposent un traitement spécifique par rapport aux documents normés produits par les professionnels de santé, une deuxième limite apparaît clairement lorsqu'il s'agit de traiter le domaine médical, d'abord pour les chercheurs en traitement automatique des langues, puis pour le TAL proprement dit : la difficulté d'accéder aux connaissances médicales. Au risque de révéler des évidences, la médecine est complexe et identifier des témoignages susceptibles d'intéresser les pharmacovigilants, ou extraire des informations utiles pour les médecins, supposent de pourvoir déterminer si les informations présentes dans une portion de texte constituent un contenu pertinent. Pendant ces travaux et parce que je n'ai pas suivi d'études médicales, j'ai souvent été confronté à la difficulté d'accéder aux connaissances médicales. S'il est relativement simple de déterminer si la valeur numérique d'un examen est normale grâce aux documentations disponibles sur internet, encore que certaines valeurs sont conditionnelles en fonction de l'âge ou du sexe, identifier les effets secondaires, puis le mésusage, et plus récemment, l'intérêt potentiel de traitements pour soigner le virus de la Covid-19, se révèle difficile pour le chercheur en TAL qui ne peut solliciter que ponctuellement l'avis de médecins. Des travaux de représentation des connaissances médicales dans des ontologies ont été réalisés dans le passé [Ceusters et al., 1998, do Amaral et al., 2000] avec la possibilité de modéliser un pourcentage élevé de connaissances dans les ontologies pour des tâches d'identification de concepts phénotypiques et génétiques [Friedman et al., 2006]. Il apparaît cependant que ce type d'approche a été remplacé par les travaux à base de plongements, comme le modèle de langue NeuroCORD [Wu et al., 2022] réalisé spécifiquement pour identifier les pathologies liées au Covid-19, ou ces modèles à base de graphe pour l'aide au diagnostic [Zhang et al., 2022], ou encore l'affinage de plusieurs modèles généraux et spécialisés (BERT, BioBERT, NCBI BERT, ClinicalBERT) pour de la découverte de connaissances sur les médicaments [Li et al., 2021]. Je présente un aperçu de cette diversité de modèles dans le chapitre suivant.

Chapitre 5
Des modèles qui suscitent des questionnements

Sommaire

5.1 Introduction

Dans cette section, et sans prétendre à une présentation exhaustive et détaillée, je dresse un panorama rapide des modèles pré-entraînés actuels dans le double objectif de fournir des clés de compréhension au lecteur et de positionner le travail de recherche réalisé avec mes doctorants parmi cette diversité.

5.1.1 Une diversité de modèles pré-entraînés

Représentation continue Les plongements de mots ont été popularisés avec Word2Vec [Mikolov et al., 2013] qui permet de prédire un mot pour un contexte donné [1], ou inversement, de prédire un contexte pour un mot donné grâce aux deux structures CBOW (continuous Bag-of-Words) et Skip-Gram. Cette première approche a été améliorée par GloVe [Pennington et al., 2014] qui repose sur une agrégation de matrices de co-occurrences de mots, puis par FastText [Bojanowski et al., 2017], une extension de Word2Vec, qui encode des sous-unités dans les vecteurs plutôt que des mots, et qui permet de couvrir 157 langues. Pour chaque mot, ces approches n'apprennent qu'un seul vecteur. On parle alors de « représentation continue » ou de plongements statiques.

Représentation contextuelle Afin de tenir compte de l'ordre des mots, les représentations contextuelles ou plongements de phrases ou plongements contextuels ont été développés, d'abord avec ELMo (Embeddings from Language Models) [Peters et al., 2018] qui a fait l'objet d'adaptation dans plusieurs langues [2]. L'architecture des *transformers* [Vaswani et al., 2017], fondée uniquement sur les mécanismes d'attention pour faire le lien entre les couches d'entrée et de sortie,

1. Dans le cadre de sa thèse, Liyun a notamment appliqué cette méthode sur le chinois (mandarin standard), au moyen du module gensim (https://pypi.org/project/gensim/) en visant deux objectifs. D'une part, produire un lexique propre à son domaine d'étude (l'hôtellerie), et d'autre part, disposer de variantes d'entités du domaine, en particulier pour la translittération de mots étrangers qui donne généralement lieu à plusieurs variantes (voir page 18). Ce travail est abordé dans son article jeune chercheur [Yan, 2018].

2. Des modèles ELMo sont disponibles en allemand, basque, japonais, et portugais pour la langue générale, ainsi qu'en anglais pour le domaine biomédical (PubMed) : https://allenai.org/allennlp/software/elmo

et surpassant en performances la structure *encoder-decoder* précédemment employée dans les approches neuronales, a permis le développement du modèle BERT (Bidirectional Encoder Representations from Transformers) [Devlin et al., 2019]. Ce premier modèle a fait l'objet d'une optimisation au niveau des hyper-paramètres et donné lieu au modèle RoBERTa [Liu et al., 2019]. Des adaptations spécifiques [3] à certaines langues ont été réalisées à partir de RoBERTa, notamment pour le français avec les modèles CamemBERT [Martin et al., 2020] (entraîné en octobre 2019 sur le corpus OSCAR [4], CCNet, et Wikipedia) et FlauBERT [5] [Le et al., 2020a, Le et al., 2020b] (entraîné en mai 2020 sur les corpus WMT19, OPUS, et Wikimedia). À date de rédaction, ces deux modèles constituent les modèles reposant sur l'architecture des transformers les plus utilisés par la communauté francophone.

Des adaptations ont également été réalisées par domaine tels astroBERT [Grèzes et al., 2021] pour le domaine astro-physique, BioBERT [Lee et al., 2019] pour le biomédical, avec une spécification accrue pour certaines tâches comme dans BioNER [Sun et al., 2021] qui permet une reconnaissance d'entités nommées en domaine biomédical, DNABERT [Ji et al., 2021] pour traiter les informations d'ADN présentes dans le génome, LegalBERT [Chalkidis et al., 2020] pour le domaine juridique, ou encore SciBERT [Beltagy et al., 2019] à partir d'un corpus d'articles scientifiques. L'affinage *(fine-tuning)*, en tant qu'adaptation au domaine, consiste à poursuivre l'entraînement d'un modèle sur les données du domaine à traiter. Il a cependant été démontré que des modèles provenant d'une poursuite d'entraînement du modèle BERT sur des corpus de domaine général ou de domaine de spécialité ou combinant les deux, obtenaient des performances similaires sur plusieurs tâches du domaine médical en anglais (repérage d'entités nommées, extraction de relations, implication textuelle) [El Boukkouri, 2020]. Des modèles ont également été entraînés depuis l'architecture d'origine BERT pour traiter des contenus particuliers, tels CodeBERT pour le code informatique [Feng et al., 2020], ou BERTweet sur des tweets rédigés en anglais pour trois tâches (classification, étiquetage en parties-du-discours, repérage d'entités nommées) [Nguyen et al., 2020].

Référentiels d'évaluation En parallèle du développement de ces modèles, des référentiels d'évaluation *(evaluation benchmark)* ont été élaborés pour plusieurs tâches du traitement automatique des langues, tels que GLUE [6] [Wang et al., 2018] et SuperGLUE [7] [Wang et al., 2019] pour l'anglais, et FLUE [8] [Le et al., 2020a] pour le français. Ce dernier propose un référentiel sur sept tâches : classification de textes, étiquetage en parties du discours, analyse syntaxique, désambiguïsation lexicale (noms et verbes), identification de paraphrases, et reconnaissance d'implications textuelles. Notons qu'il existe également des référentiels par domaine, notamment BLUE [9] [Peng et al., 2019] pour le biomédical sur cinq tâches en anglais (reconnaissance d'entités nommées, extraction de relations, inférences textuelles, similarité textuelle, classification multi-labels) et BLURB [10] [Gu et al., 2022] sur six tâches en biomédical (reconnaissance d'entités nommées, extraction de relations, similarité textuelle, classification de documents, question-réponse, et extraction d'informations).

3. Voir https://huggingface.co/docs/transformers/index pour une liste des modèles transformers disponibles pour les librairies PyTorch, TensorFlow, et JAX. À date de rédaction, plus d'une centaine de modèles apparaît dans cette liste, soit dédiés à des langues (allemand, français, multilingue), soit à des formats de données (CSV, XML), soit encore pour certaines tâches (recherche d'information [Reimers and Gurevych, 2019], reconnaissance de la parole, transcription de la parole, traduction).

4. OSCAR : Open Super-large Crawled Aggregated coRpus, corpus multilingue composé de pages web rédigées dans 151 langues (version 22.01), https://huggingface.co/oscar-corpus

5. FlauBERT : French Language Understanding via Bidirectional Encoder Representations from Transformers.

6. GLUE : General Language Understanding Evaluation, https://gluebenchmark.com/

7. SuperGLUE, https://super.gluebenchmark.com

8. FLUE : French Language Understanding Evaluation, https://github.com/getalp/Flaubert/tree/master/flue

9. BLUE : Biomedical Language Understanding Evaluation, https://github.com/ncbi-nlp/BLUE_Benchmark

10. BLURB : Biomedical Language Understanding and Reasoning Benchmark, https://microsoft.github.io/BLURB/

5.1.2 Des modèles coûteux à produire et à utiliser

Coût à l'utilisation : des modèles allégés Parce que ces modèles encodent un nombre élevé d'informations et qu'ils permettent de traiter plusieurs tâches, des versions allégées ont été produites à partir de l'architecture originale BERT pour réduire le coût à l'utilisation. Des solutions reposant sur des techniques de compression ont notamment été proposées. Les auteurs de ces solutions assurent du maintien des performances sur le référentiel GLUE par rapport au modèle d'origine. Une première version a été produite avec le modèle ALBERT (dont le nom fait écho à la description *A Lite BERT*) [Lan et al., 2019], en décomposant la matrice du vocabulaire en deux matrices plus petites. Une version entraînée sur des données Wikipedia a été produite pour le français en septembre 2021, FrALBERT [Cattan et al., 2021], pour la génération de phrases dans les systèmes de question-réponse. Pour autant que nous puissions en juger et contrairement à ce que pourrait laisser croire le nom, le modèle italien AlBERTo [11] [Polignano et al., 2019] ne dérive pas du modèle ALBERT mais directement du modèle original BERT. Une deuxième optimisation du modèle original BERT a été réalisée avec le modèle BART (Bidirectional and Auto-Regressive Transformers) [Lewis et al., 2020], entraîné avec une fonction de bruit sur une version corrompue du texte pour reconstruire le texte d'origine, puis adapté au français en octobre 2020 avec le modèle BARThez [Kamal Eddine et al., 2021], actuellement le quatrième modèle de langue disponible pour le français. Enfin, bien qu'entraîné sur les mêmes corpus que l'architecture de base BERT (Wikipedia en anglais et le Toronto Book Corpus), le modèle DistilBERT [Sanh et al., 2019] conserve 97% des informations sémantiques du modèle d'origine mais permet une utilisation plus rapide puisque la taille est réduite de 40% grâce à une technique de compression appelée « distillation de connaissances » [Hinton et al., 2014]. Cette distillation a été appliquée par des apprentistes en juillet 2022 sur le modèle CamemBERT pour produire le cinquième modèle disponible en français, nommé DistilCamemBERT [Delestre and Amar, 2022].

Coût à l'entraînement : pallier un réentraînement coûteux Face aux coûts financiers et environnementaux élevés d'un entraînement de plongements contextuels [Strubell et al., 2019], y compris pour de l'affinage par rétro-propagation, tant pour l'adaptation au domaine que l'adaptation aux données, des approches consistant à combiner aux vecteurs du modèle d'origine les vecteurs appris sur des textes du domaine ou d'une tâche par combinaison linéaire ont été proposées [Labutov and Lipson, 2013]. Dans le domaine biomédical, nous observons ces dernières années une tendance consistant à réemployer les architectures « anciennes » telles que Word2Vec et FastText, pour encoder les informations de base (essentiellement les classes lexicales, orthographiques, et syntactico-sémantiques) en vue de les réinjecter dans les vecteurs du modèle de langue d'origine *(vector retrofitting)* [Grabar and Grouin, 2022]. C'est notamment l'approche qui a été suivie pour constituer le modèle GreenBioBERT [Poerner et al., 2020], en utilisant Word2Vec sur des textes du domaine (les articles scientifiques issus de PubMed et PubMed Central) et en alignant les vecteurs obtenus avec ceux du modèle BERT. Sur du repérage d'entités nommées, leur modèle obtient de meilleures performances que le modèle BERT d'origine, et des performances modérément plus faibles quoi que encore compétitives par rapport aux versions disponibles du modèle BioBERT [Lee et al., 2019]. Il a également été démontré que des plongements statiques de langue générale produits par ELMo suffisent à enrichir des plongements appris sur un petit volume de données du domaine médical [El Boukkouri et al., 2019] par opposition à des plongements appris sur un gros volume de données du domaine.

Le tableau 5.1 propose une synthèse non exhaustive des versions dérivées des modèles de représentations contextuelles d'origine ELMo et BERT, en terme d'optimisations algorithmiques (hyper-paramètres, version allégée, encodage d'une fonction de bruit), d'adaptation au domaine ou en langue issue de ces optimisations, et des améliorations méthodologiques apportées aux adaptations réalisées. Ces modèles ont été produits en l'espace de quatre ans seulement.

11. AlBERTo : https://github.com/marcopoli/AlBERTo-it

Base	Optimisation	Adaptation	Amélioration méthodologique
ELMo (2018)		ELMo (médical)	
		ELMo (de, eus, jp, pt)	
BERT (2019)		astroBERT (astrophysique, 2022)	
		BioBERT (biomédical, 2019)	
			GreenBioBERT (2020)
		DNABERT (génome, 2021)	
		LegalBERT (légal, 2020)	
		SciBERT (scientifique, 2019)	
		BERTweet (tweet, 2020)	
		CodeBERT (code, 2020)	
		SentenceBERT (2019)	
	ALBERT (2019)	FrALBERT (fr, 2021)	
	BART (2020)	BARThez (fr, 2021)	
	CharacterBERT (2020)		
	DistilBERT (2019)		DistilCamemBERT (fr, 2022)
	RoBERTa (2019)	CamemBERT (fr, 2019)	CamemBERT-POS (2021, p. 75)
		FlauBERT (fr, 2020)	FlauBERT-POS (2021, p. 75)

TABLE 5.1 : Optimisations algorithmiques, adaptations au domaine ou en langue, et améliorations méthodologiques dérivées des modèles de représentations contextuelles de base ELMo et BERT

5.1.3 Problématique de la représentation des textes

La représentation des textes dans les modèles de langue passe par une segmentation en sous-unités, parmi trois méthodes principales : l'algorithme de compression de données BPE (Byte-Pair Encoding) [Gage, 1994, Sennrich et al., 2016], le segmenteur en tokens *(tokenizer)* WordPiece [Schuster and Nakajima, 2012], ou encore l'outil Unigram [Kudo and Richardson, 2018]. De manière à réduire la complexité algorithmique, les sous-unités créées par les modèles transformers existent en nombre limité (au maximum 50.000 entrées [12], généralement moins). Ces sous-unités sont créées de manière statistique et ne correspondent, ni à des racines linguistiques, ni à des morphèmes de la langue, même s'il n'est pas exclu qu'elles coïncident avec des composants linguistiques. Cette problématique de tokénisation n'est pas sans conséquence pour le TAL. Elle implique également une problématique inverse de retokénisation une fois les traitements terminés.

Mots hors vocabulaire (HV)

Cette problématique se trouve renforcée avec les mots hors-vocabulaire qui ne sont pas rares en langue, et qui ont un impact sur les performances des systèmes développés (reconnaissance et compréhension de la parole, repérage d'entités nommées). Par définition, il s'agit de séquences de caractères qui sont absentes du dictionnaire utilisé. Cette absence peut s'expliquer, soit par la taille limitée du vocabulaire qui exclut les entités nommées de type noms propres, soit par l'utilisation d'un lexique de langue générale pour traiter une langue de spécialité, ou encore par l'apparition de néologismes ou de nouveaux usages (code switching, français inclusif, spécificité des réseaux sociaux, etc.) qui n'ont pas encore été intégrés dans les dictionnaires employés (telles que les abréviations dans les salutations : « *slt* », « *cdlt* »). On parle également de mots hors vocabulaire dans le cas d'erreurs commises par l'utilisateur, comme dans la production d'un texte désaccentué [Zweigenbaum and Grabar, 2002], ou par les systèmes tels que les systèmes de reconnaissance optique de caractères *(Optical Character Recognition)* qui peuvent générer une séquence de caractères absente du dictionnaire et nécessiter un traitement dédié [Oprean et al., 2014].

12. https://huggingface.co/docs/transformers/tokenizer_summary

Sur un domaine de spécialité, et parce que les sous-unités de langue générale produites par le segmenteur WordPiece ne correspondent qu'imparfaitement à la terminologie du domaine médical, des expériences ont été menées pour représenter les mots par caractères, en utilisant un réseau de neurones de type Character-CNN sur le corpus clinique MIMIC-III [Johnson et al., 2016]. Si les expériences réalisées permettent une amélioration des résultats avec des représentations robustes [El Boukkouri, 2021, El Boukkouri et al., 2021], les auteurs ajoutent que la production de plongements lexicaux sur des bases de connaissances pertinentes du domaine est à considérer en complément de cette approche.

5.2 Impact de la segmentation sur les mots hors vocabulaire

Introduction Nous avons étudié pendant la thèse d'Alexandra Benamar l'impact des tokéniseurs sur la représentation des mots HV en français, en comparant les modèles CamemBERT (fondé sur le tokéniseur SentencePiece) et FlauBERT (algorithme de compression BPE) sur trois types de mots HV : les termes d'un domaine de spécialité (la microbiologie : « *eucaryote* », « *protozoaire* », etc.), les fautes d'orthographe (« **infractus* » au lieu d'*infarctus*), et les homographes entre langues de spécialité (polysémie du terme « *filiation* » selon qu'il s'agit du domaine juridique ou biologique). Pour chacun des modèles CamemBERT et FlauBERT, des variantes ont été proposées (concaténation avec ELMo pour ajouter des informations contextuelles ; affinage sur les données traitées pour apprendre les mots HV ; et ajout d'informations de parties du discours de manière alternative à l'affinage, voir section 5.4). Les corpus ont été choisis pour les différents types de mots HV qu'ils contiennent : un corpus d'articles scientifiques numérisés du *Journal de Microbiologie* parus entre 1887 et 1900 disponible sur la librairie numérique Gallica et caractérisé par des erreurs d'OCR ; un corpus d'articles de loi utilisé pour la campagne DEFT 2006 [Azé et al., 2006] et composé de plusieurs mots désaccentués ; et un corpus privé d'échanges par mails entre la clientèle et EDF contenant des fautes d'orthographe, des abréviations, et des noms de produits commerciaux.

Le tableau 5.2 présente le résultat de la tokénisation en sous-unités pour quatre termes utilisés dans le corpus de mails EDF (« *compteur* », « *cordialement* », « *linky* », « *remboursement* »), en fonction du corpus utilisé pour entraîner le modèle CamemBERT. Nous observons que le modèle entraîné sur Wikipédia n'est pas suffisamment performant pour traiter les termes utilisés dans un corpus d'échanges électroniques caractérisé par des formules de salutation (« *cordialement* »), ni pour traiter les termes d'un domaine de spécialité (« *compteur* », « *linky* » et « *remboursement* »), alors que les corpus issus de contenus internet (CCNet et OSCAR) permettent de conserver les termes de la langue (seul le produit commercial est tokénisé en plusieurs sous-unités).

Terme	Wikipédia (base)	CCNet (base/large), OSCAR (base)
« *compteur* »	[“_compte”, “ur”]	[“_compteur”]
« *cordialement* »	[“_cord”, “iale”, “ment”]	[“_cordialement”]
« *linky* »	[“_l”, “in”, “ky”]	[“_l”, “ink”, “y”]
« *remboursement* »	[“_rem”, “bour”, “s”, “ement”]	[“_remboursement”]

Table 5.2 : Exemple de tokénisations en sous-unités sur quatre termes du corpus d'échanges électroniques EDF, en fonction du corpus utilisé pour entraîner le modèle CamemBERT

5.2.1 Analyse de trois types de mots hors vocabulaire

Termes de domaines de spécialité

Les expériences des termes du domaine de spécialité reposent sur une dizaine de mots HV parmi les plus fréquents dans les corpus Gallica et DEFT [13], et sur le calcul du coefficient de Dice (formule 5.1 avec n_X et n_Y le nombre de n-grammes dans X et Y et n_Z le nombre de n-grammes communs entre X et Y), et d'une variante proposée par Alexandra nommée Dice-SU, fondée sur les sous-unités générées pendant la tokénisation plutôt que les n-grammes (formule 5.2 avec $t_M(X)$ et $t_M(Y)$ la fonction de tokénisation utilisée sur X et Y avec le modèle M), entre les mots HV et leurs cinq plus proches voisins par similarité cosinus.

$$\text{Dice(X,Y)} = 2 \times \frac{n_Z}{n_X + n_Y} \qquad (5.1)$$

$$\text{Dice-SU(X,Y)} = 2 \times \frac{n_{t_M(Z)}}{n_{t_M(X)} + n_{t_M(Y)}} \qquad (5.2)$$

La figure 5.1 présente la distribution des coefficients Dice et Dice-SU des mots HV par modèle et variante sur les corpus Gallica et DEFT. Nous observons que la valeur du coefficient de Dice

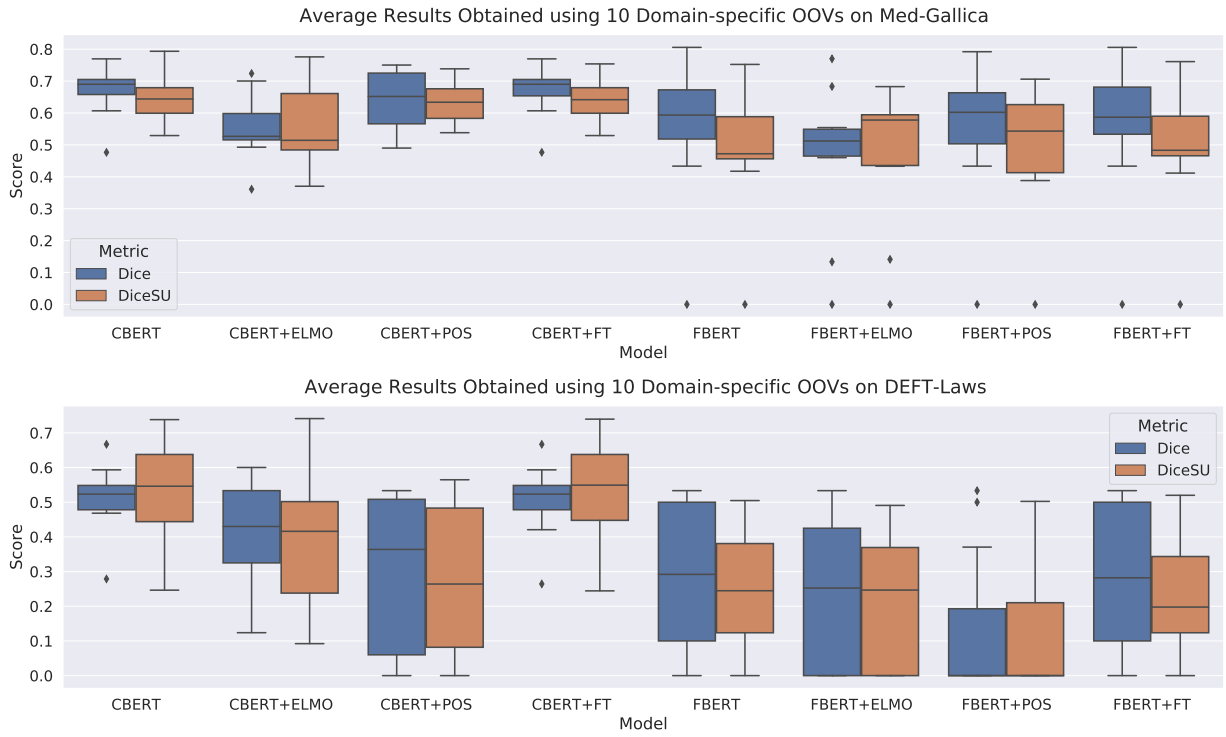

FIGURE 5.1 : Distribution des coefficients Dice (bleu) et Dice-SU (orange) des mots HV par modèle (CamemBERT et FlauBERT) et variante sur les corpus Gallica (haut) et DEFT (bas)

est en moyenne plus élevée avec CamemBERT qu'avec FlauBERT, ce qui suggère que les mots HV sont davantage affectés par la représentation utilisée par CamemBERT. Avec le coefficient de Dice calculé sur les sous-unités (Dice-SU), les valeurs sont faibles sur le corpus DEFT, suggérant que les voisins distributionnels ne partagent que très peu de sous-unités, à l'inverse de ce qui est observé sur Gallica, avec 50 à 70% de sous-unités partagées entre les mots HV et leurs voisins.

13. Mots HV les plus fréquents dans Gallica *(incubation, bactériologique, épileptique, prophylactique, tuberculose, cautérisation, bacillophage, sepsis, hyperesthésie, anorexie)* et DEFT *(allégation, frauduleux, minutes des actes, délibéré, loi, apparance, discriminatoire, régularisé, cessionnaire, national, régularisé, déposé, scellé, apposé).*

Fautes de frappe et erreurs orthographiques

L'analyse de l'impact de ces fautes repose sur des types d'erreurs propres à chaque corpus : des erreurs sur des mots du vocabulaire de CamemBERT et FlauBERT pour le corpus DEFT (« **xonditions* » au lieu de « *conditions* »), les erreurs sur des mots du domaine absents du vocabulaire pour le corpus Gallica (« **injoction* » au lieu de « *injection* »), et les erreurs sur les termes propres aux échanges formels du corpus de mails (« **cordiallement* » au lieu de « *cordialement* »). Cent mots erronés ont été identifiés par corpus, puis le coefficient de Dice a été calculé entre versions erronée et correcte. Le tableau 5.3 présente la moyenne des valeurs de similarité cosinus obtenues sur les cent mots HV de chaque corpus, en fonction du modèle et de la variante utilisée.

Corpus	CamemBERT			FlauBERT		
	base	+ELMo	+POS	base	+ELMo	+POS
DEFT	0,19	0,32	**0,63**	0,15	0,34	**0,56**
Gallica	0,39	0,54	**0,66**	0,37	0,57	**0,63**
Mails EDF	0,27	0,44	**0,92**	0,34	0,56	**0,93**

TABLE 5.3 : Moyenne des similarité cosinus entre les versions erronée et correcte des mots HV

Nous observons sur chaque corpus que la similarité est plus élevée avec la variante de CamemBERT ou de FlauBERT intégrant les informations de parties du discours dans les plongements, suggérant que les informations morpho-syntaxiques fournissent une information contextuelle utile pour le traitement des mots HV. Sur le corpus Gallica, les résultats plus élevés semblent indiquer que les modèles capturent plus facilement des informations sémantiques de proximité autour des mots HV contenant des erreurs d'OCR. Sur le corpus d'échanges électroniques, les résultats élevés s'expliquent également par la position des erreurs dans le mot, comme suggéré dans d'autres études où les erreurs en début de mots affectent davantage les performances des tokéniseurs de modèles [Nayak et al., 2020]. Alexandra a étudié ce phénomène sur deux mots fréquents dans le corpus de mails (« *cordialement* » et « *remboursement* ») en étudiant la tokénisation réalisée selon que l'erreur est au début, au milieu, ou en fin de mot. Sur le terme « *cordialement* », alors que la version correcte est tokénisée en une seule unité *_cordialement* par CamemBERT (pour la version entraînée sur OSCAR), la version erronée en début de mot « *ccordialement* » sera tokénisée en quatre sous-unités *_c,cord,iale,ment* [Benamar et al., 2021]. Une analyse des similarités cosinus calculées entre les formes erronée et correcte confirme la plus grande difficulté à traiter les erreurs en début de mot (figure 5.2).

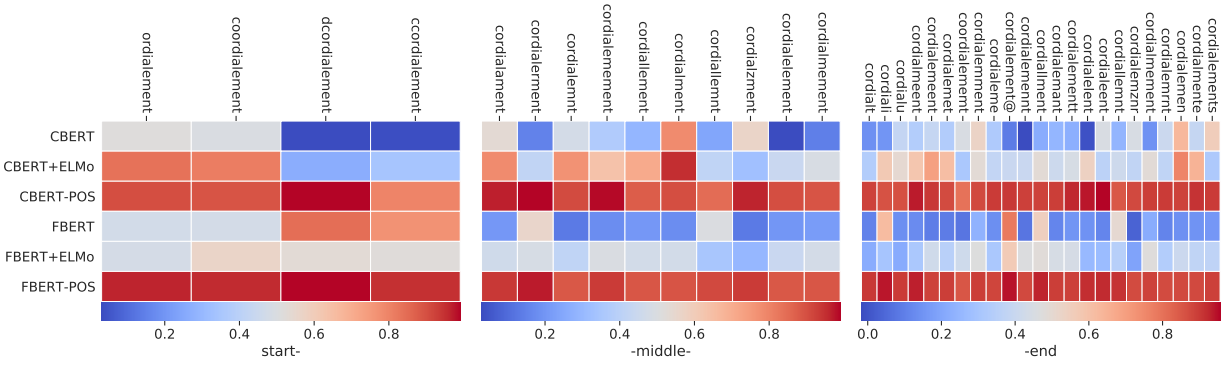

FIGURE 5.2 : Similarité cosinus entre formes erronée et correcte selon que l'erreur est en début de mot (gauche), au milieu (centre), ou en fin de mot (droite)

Homographes

Enfin, l'analyse des homographes a porté sur les termes polysémiques ayant une signification différente entre deux langues de spécialité (juridique dans DEFT vs. biomédical dans Gallica) parmi quatre termes polysémiques (« *preuve* », « *filiation* », « *observation* », et « *isolement* »). L'évaluation a porté sur le pourcentage de mots cooccurrant dans le voisinage de chaque homographe, parmi les dix plus proches voisins, entre les corpus DEFT et Gallica. Les résultats obtenus ne mettent pas en avant de termes spécifiques à l'un des domaines de spécialité traités (avec une similarité contextuelle qui monte à 90% pour le mot « *preuve* » avec CamemBERT).

5.2.2 Une représentation du texte qui impacte les tâches de TAL

Dans ce travail sur l'impact de la tokénisation [14] par des modèles CamemBERT et FlauBERT, nous relevons plusieurs points [Benamar et al., 2022c]. Une première conclusion concerne le traitement des erreurs produites par un OCR (corpus Gallica) qui donne de meilleurs résultats que celui des fautes de frappe des utilisateurs (corpus DEFT), alors même que les mots corrects étaient présents dans le vocabulaire des modèles testés pour les erreurs du corpus DEFT et qu'inversement, les mots corrects du domaine de spécialité étaient absents. Une deuxième conclusion porte sur les variantes de modèles employés. Nous validons la solution d'enrichissement des plongements par l'ajout des parties du discours, et sur laquelle je reviendrai en section 5.4, à préférer à l'affinage sur les données du corpus pour apprendre les mots HV. Une dernière conclusion concerne l'impact de la position des erreurs sur les performances de tokénisation des modèles transformers. Nous confirmons l'hypothèse selon laquelle des erreurs en début de mot ont un impact plus fort sur la tokénisation que des erreurs situées en milieu ou en fin de mot.

5.3 Des modèles qui n'encodent pas que la sémantique

5.3.1 Identification de stéréotypes dans des productions textuelles

Les modèles actuels sont entraînés sur des corpus de documents, pour capturer la sémantique contenue dans ces documents [Sahlgren, 2008] (voir section 1.2), dans l'objectif de s'appuyer sur cette sémantique pour diverses tâches de TAL [El Boukkouri, 2020] et divers domaines spécialisés [Lee et al., 2019, Grèzes et al., 2021] (voir section 5.1). Plusieurs études ont montré que les modèles n'encodent pas que la sémantique, mais également des « biais » qui témoignent de la représentation mentale des locuteurs sur leur société. Dans ses travaux de thèse sur l'adaptation au domaine, Morgane Marchand a mis en évidence l'existence de ces biais et s'est appuyée sur ces informations pour une tâche de fouille d'opinion [Marchand, 2015]. D'autres études ont mis en évidence les biais sociétaux contenus dans ces modèles, comme les biais de genre [Sun et al., 2019], présents dans les messages sur Twitter [Park et al., 2018]. Notons qu'un workshop scientifique portant spécifiquement sur la question des biais de genre en TAL [15] se tient tous les ans depuis 2019. Des corpus ont également été constitués dans plusieurs langues pour illustrer ces biais. Le corpus anglais *CrowS-pairs* [Nangia et al., 2020] se compose de 1508 paires de phrases représentant neuf types de biais (âge, apparence physique, ethnie et couleur de peau, genre, nationalité, orientation sexuelle, pathologie, religion, et statut socio-économique). Une traduction de ce corpus intitulée *French CrowS-Pairs* a été réalisée en français et complétée de paires de phrases plus représentatives de la société française [Névéol et al., 2022]. Il est cependant légitime de s'interroger sur l'utilité de tels corpus qui concentrent des exemples qu'il est possible de qualifier de « clichés » [16].

14. Le code informatique produit par Alexandra dans le cadre de cette étude est disponible sur son dépôt GitHub : https://github.com/alexandrabenamar/evaluating_tokenizers_oov

15. Workshop GeBNLP (Gender Bias in Natural Language Processing) : https://aclanthology.org/venues/gebnlp/

16. Les *{vieux, gros, blancs, hommes, français, hétéros, cancéreux, chrétiens, pauvres}* sont *ADJECTIF QUALIFICATIF DÉPRÉCIATIF* / ne savent pas *VERBE* / ont *EXPRESSION GÉNÉRALISANTE*. Par exemple, « *Les vieux sont lents.* », « *Les hommes ne savent pas faire la vaisselle.* », « *Les gros ont horreur du sport.* ».

Je considèrerai ces ressources comme permettant a minima une prise de conscience du phénomène, dans sa diversité des biais existants, mais offrant également des exemples dont on peut s'inspirer pour révéler les biais présents dans les modèles produits.

Le concept même de « biais » a finalement été assez peu questionné, en dehors d'une critique assez vive qui relève l'incohérence d'étudier un phénomène qui n'a pas été correctement défini (avec des définitions parfois différentes entre auteurs) et l'absence de correspondance entre résultats obtenus et motivations mises en avant pour travailler sur ces biais (aucun article n'explique en quoi ces biais sont dommageables, dans quel sens, et auprès de qui) [Blodgett et al., 2020]. On relativisera toutefois cette observation suite à l'expérience désastreuse de *chatbot* @Tay développé par Microsoft Research en 2016 sur le réseau Twitter, et dont le compte a été fermé au bout d'une journée seulement et 96 000 tweets générés [17], après que les utilisateurs du réseau aient réussi à faire tenir des propos racistes et misogynes par ce chatbot. Ces propos ont toutefois été rendus possibles par l'entraînement du chatbot, en partie sur des éléments de répartie produits par des comédiens spécialistes de l'improvisation, et par sa capacité à répéter les accroches formulées par les utilisateurs du réseau [Neff and Nagy, 2016]. Par ailleurs, nous pouvons nous interroger si le problème réside réellement dans les données d'entrées (que l'on pourrait considérer par analogie comme le cerveau contenant les connaissances sur le monde et la langue), ou dans l'absence de réflexion sur ce qui est produit en sortie (à l'image d'une personne qui réfléchit en même temps qu'elle exprime cette idée, donc sans avoir de recul sur les propos tenus, et sans activer de filtre au moment de l'oralisation). Lorsque des motivations existent, elles révèlent ainsi l'inquiétude des conséquences négatives qui pourraient être faites des modèles intégrant ces biais, quand bien même ces conséquences n'ont pas été mesurées. Parmi les solutions envisagées figure le fait d'écarter les contenus négatifs, pour éviter un entraînement sur ce type de données [Bridge et al., 2021].

5.3.2 Application au théâtre classique français

Au début de la thèse d'Alexandra Benamar, et dans un objectif de découverte des plongements lexicaux avec l'outil word2vec, nous avons étudié les stéréotypes [18] présents dans les représentations vectorielles, sur un corpus de pièces de théâtre classiques [Benamar et al., 2022b]. Au-delà de l'identification des stéréotypes, et même s'il n'était pas prévu de réutiliser les vecteurs produits, une vertu de ce travail reste la sensibilisation à ce phénomène en prévision de futurs travaux liés à la manipulation de données sensibles (telles que les communications de la clientèle EDF). Nous avons choisi d'axer cette recherche sur des pièces de théâtre françaises, parues entre le XVI^{e} et le XIX^{e} siècle. D'une part, en raison de la disponibilité de ces pièces depuis le site https://www.theatre-classique.fr/ en licence Creative Commons BY-NC-ND, d'autre part, parce que l'écriture du théâtre amplifie les stéréotypes existant dans le monde réel à des fins de critique sociale ou de comédie [Marcandier, 2011]. Nous focalisons notre recherche sur deux genres théâtraux (la comédie et la tragédie) en étudiant trois dimensions : la sémantique des voisins distributionnels des termes « *femme* » et « *homme* », l'association des émotions au genre, et les voisins distributionnels de six personnages types du théâtre classique.

Stéréotypes de genre en fonction du corpus d'entraînement

Une première analyse des voisins distributionnels des termes « *femme* » et « *homme* », dans des modèles existants entraînés sur Wikipedia ou internet [19] et dans des modèles entraînés conjointement ou séparément sur la comédie et la tragédie, montre que les voisins (par la distance cosinus)

17. Le compte @Tay et le site web associé ont été fermés, et renommés @Zo en fin d'année 2016 en raison de l'image désastreuse pour la société Microsoft. Il semble néanmoins que cette nouvelle version ait été elle aussi capable de révélations fracassantes, comme le fait que le système d'exploitation Windows serait un logiciel espion (https://www.insider.com/microsoft-ai-chatbot-zo-windows-spyware-tay-2017-7). Ce chatbot a fermé en mars 2019.
18. En raison de la difficulté à définir ce qu'est un biais, nous décidons de travailler sur le concept de « stéréotype ».
19. Modèles produits par Jean-Philippe Fauconnier, disponibles depuis https://fauconnier.github.io/

ne renvoient pas aux mêmes propriétés (voir tableau 5.4) : pour les femmes, les plus proches voisins renvoient aux rapports sociaux entretenus avec les hommes *(épousé, fille, maîtresse, veuve, fiancée)*, à l'âge ou au physique *(jeune, adolescente, beauté)*, et à la sexualité *(prostituée, maîtresse, vertueuse)*, tandis que pour les hommes, les voisins renvoient essentiellement à des traits de caractère qui ne sont pas nécessairement positifs *(coureur, joueur, rustre, bravoure)*.

Terme	Modèles existants		Modèles entraînés sur le théâtre classique		
	Wikipedia	Internet	Comédie + Tragédie	Comédie	Tragédie
femme	fille, maîtresse, prostituée, fiancée, servante	fille, mari, jeune, prostituée, adolescente	mari, belle-sœur, vindicative, accoucha, veuve	mari, veuve, feue, épousé, vertueuse	mari, veuve, ressource, belle-mère, comtesse
homme	diplomate, femme, politicien, avocat, agriculteur	humaine, galop, humanité, nationalisme, sabre	garçon, hébété, désintéressé, aventurier, dissipateur	rustre, coureur, moine, joueur, bravoure	méchant, fou, franc, président, raisonnable

TABLE 5.4 : Cinq plus proches voisins des termes « femme » et « homme » dans des plongements appris sur Wikipedia, internet, et des pièces de comédie et tragédie du théâtre classique

Le genre des émotions en fonction du corpus d'entraînement

Une deuxième analyse porte sur les quatorze émotions définies dans une étude sur le genre des émotions [Raymondie and Steiner, 2020]. La figure 5.3 présente le résultat de la similarité cosinus calculée entre chaque émotion et les termes « femme » et « homme », sur trois modèles word2vec (les modèles existants entraînés sur Wikipedia et internet, et celui entraîné sur le théâtre).

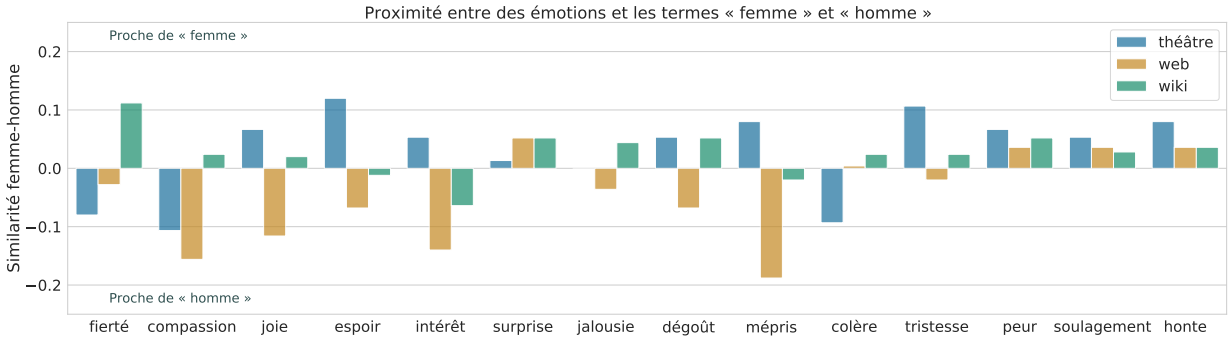

FIGURE 5.3 : Proximité par la similarité cosinus entre émotions et termes « femme » et « homme »

La proximité entre le vecteur d'émotion $v_{\text{émo}}$ et les vecteurs femme v_f et homme v_h a été calculée au moyen de la formule 5.3. Une valeur mathématiquement positive correspond à une émotion plutôt féminine tandis qu'une valeur négative correspond à une émotion plutôt masculine.

$$Proximité = cos(v_{\text{émo}}, v_f) - cos(v_{\text{émo}}, v_h) \tag{5.3}$$

Si quatre émotions apparaissent majoritairement féminines dans les trois modèles comparés *(honte, peur, soulagement, surprise)*, nous observons que le genre associé aux autres émotions varie en fonction des modèles contextuels étudiés. Nous constatons que les différences sont plus marquées sur le théâtre classique (en bleu), avec des valeurs de similarité cosinus plus éloignées du zéro, sauf pour deux émotions qui se révèlent masculines sur le web *(intérêt, mépris)*.

Stéréotypes autour des personnages types du théâtre classique

Une dernière analyse concerne les voisins distributionnels de six personnages types du théâtre classique, avec un rôle non-genré *(enfant)* et cinq rôles antagonistes (*roi* et *reine*, *valet* et *servante*, et *maître* par une opposition de classe avec le valet), sur les modèles issus du corpus de théâtre, selon que les termes sont employés dans une comédie ou dans une tragédie. Le tableau 5.5 présente les cinq plus proches voisins de ces personnages dans les deux genres théâtraux retenus.

Personnage	Comédie	Tragédie
roi	prince, successeur, monarque, empereur, ambassadeurs	prince, nom, couronne, choix, droits
reine	souveraine, héritière, prisonnière, orgueilleuse, ingrate	faveur, honteux, princesse, hymen, jaloux
maître	brave, scélérat, instruit, successeur, imposteur	valet, laquais, riche, province, méchant
valet	coquin, vilain, fou, fripon, drôle	fripon, faquin, impertinent, domestique, arlequin
servante	demoiselle, andouille, soubrette, impertinente, gueuse	commère, minette, chonchon, chienne, hymen
enfant	bâtard, riche, jeune, mari, femme	veuve, gendre, comte, nièce, orphelin

TABLE 5.5 : Cinq plus proches voisins (distance cosinus) de personnages types (modèles pré-appris et entraînés sur le corpus de théâtre) pour les genres comédie et tragédie

Nous observons que les différences sont réelles selon le genre théâtral, avec des voisins distributionnels qui révèlent les principaux ressorts du théâtre classique pour chaque personnage type. Les personnages féminins se voient associés à l'*hymen* dans les tragédies, soit dans l'objectif d'un héritier (reine), soit dans le cas d'un bâtard (servante). Le roi est un personnage de haut-rang, avec des responsabilités et qui se cherche un successeur, tandis que la reine est plutôt associée à des valeurs négatives, comme l'est également la servante. Le valet est un personnage-clé qui semble jouer le même rôle dans les comédies et les tragédies, c'est un *fripon* et un *coquin* sur lequel on ne saurait compter. Le maître est un bourgeois, il est *riche* et *instruit* et se voit associé aux stéréotypes de cette classe sociale *(imposteur, méchant)*. Enfin, concernant le seul personnage non-genré, du fait de son jeune âge, la sémantique de l'enfant est associée à celle de la famille.

Dans ce travail d'identification des stéréotypes dans les modèles word2vec du théâtre classique, par opposition aux modèles entraînés sur Wikipedia ou sur internet, nous observons que chaque modèle encode des stéréotypes qu'il est possible de relier au genre ou au rôle dans le cas des personnages types du théâtre. Nous constatons toutefois que les voisins distributionnels et les émotions révélées varient en fonction du corpus qui a servi à entraîner les modèles.

5.4 Une alternative au réentraînement et à l'affinage

5.4.1 Ajout d'informations morpho-syntaxiques aux représentations

Comme je l'indiquais en introduction de ce chapitre (section 5.1.2), les modèles de langue peuvent être relativement coûteux à utiliser en fonction du nombre de caractéristiques encodées. Ils sont également moins performants sur des productions langagières issues d'échanges électroniques ou pour traiter des termes d'un domaine de spécialité. Dans le but d'appliquer les modèles existants en français sur des documents composés de mots absents du vocabulaire utilisé par ces modèles (voir section 5.2 sur la segmentation des mots hors vocabulaire), nous avons défini et appliqué une méthode alternative au réentraînement et à l'affinage que nous estimons relativement simple et suffisamment générique pour être déclinée sur d'autres langues et pour d'autres tâches.

L'approche envisagée par Alexandra Benamar [Benamar et al., 2021] encode des informations morpho-syntaxiques dans des plongements de type BERT. Notre hypothèse est que l'ajout d'informations morpho-syntaxiques apporte une information utile pour les modèles dans le traitement des mots HV ou dans le cas des langues de spécialité, et que ces informations sont suffisantes pour éviter le coût d'un réentraînement du modèle, ou d'une adaptation aux données par affinage. Les expériences menées reposent sur CamemBERT (fondé sur le segmenteur SentencePiece) et sur FlauBERT (fondé sur l'algorithme BPE), en raison de la possibilité de reconstruire les mots à partir des sous-unités produites par ces deux modèles. La méthode proposée reconstruit chaque mot par retokénisation et associe la partie du discours correspondante en utilisant le caractère souligné comme séparateur. La séquence produite est de la forme « $mot_1\_pos_a$, $mot_2\_pos_b$, $mot_3\_pos_a$, [...], $mot_n\_pos_m$ ». Bien que la désambiguïsation permise par l'ajout des informations morpho-syntaxiques ne soit pas pertinente pour les modèles contextuels, nous inscrivons ce travail dans un contexte de solution alternative aux approches coûteuses généralement utilisées pour traiter de nouvelles données.

Évaluation Une étude des voisins distributionnels des termes « *cordialement* » (adverbe), « *linky* » (nom propre), et « *remboursement* » (substantif) a été effectuée, en comparant le corpus d'entraînement (OSCAR vs. Wikipédia) et la version du modèle CamemBERT parmi trois versions (CamemBERT, CamemBERT-POS, ou l'affinage de CamemBERT). Le tableau 5.6 présente les cinq plus proches voisins identifiés pour chacun des trois termes précédents.

Terme cherché	Version	Entraînement Wikipédia	Entraînement OSCAR
Cordialement (adverbe)	CamemBERT	cordialement, *cordialment, *cordialemment, cordiales, *cordialementt	merci, bonne, ph, obtenez, sincère
	CamBert-PoS	*cordiallement, cordiales, franchement, amicalement, chaleureusement	*cordiallement, chaleureusement, sincèrement, infiniment, remerciant
	Affinage	*cordialment, cordiales, *cordiallement, cordiale, *cordialelent	restant, si, merci, quelle, bonne
Linky (nom propre)	CamemBERT	linki, linkin, linke, linkey, linké	linkys, linkie, linké, linked, linkl
	CamBert-PoS	linki, ld, li, link, log	ginko, zac, cbe, log, installateur
	Affinage	linki, lindky, linly, lynky, linxy	linkie, linked, linkdy, linké, linkys
Remboursement (substantif)	CamemBERT	remboursements, *remboursment, *rembourssement, remboursés, remboursable	règlement, débit, transfert, retrait, rétablissement
	CamBert-PoS	remboursements, *rembousementt, reglement, règlement, régularisations	services, intervention, règlement, télépaiement, besoin
	Affinage	remboursements, *remboursemment, *remboursrment, *remboursemenr, *reimboursement	règlement, informée, surtout, non, gratuit

TABLE 5.6 : Cinq plus proches voisins en fonction du corpus d'entraînement de CamemBERT et de la version utilisée (CamemBERT, CamemBERT-POS, affinage CamemBERT)

Conclusions Une première conclusion montre que le modèle entraîné sur Wikipédia, quelle que soit la version, propose des variantes orthographiques y compris des erreurs ou des fautes de frappe, généralement fondées sur la même racine (« *remboursable* », « *remboursements* », « **remboursemment* ») tandis que le modèle entraîné sur OSCAR propose des variantes sémantiques (« *règlement* », « *télépaiement* », « *transfert* »). Une deuxième conclusion montre que l'approche CamemBERT-POS se fonde sur les parties du discours pour proposer des termes appartenant à la même catégorie morpho-syntaxique : les substantifs *intervention, règlement, télépaiement* pour le terme « *remboursement* », les adverbes *chaleureusement, infiniment, sincèrement* pour « *cordialement* », et les noms propres *ginko, cbe* pour le nom de produit « *linky* » (Ginko est le système d'information Enedis fondé sur le compteur Linky, CBE désigne le compteur bleu électrique). Les candidats proposés sur les noms de produits commerciaux par l'approche d'Alexandra sont ainsi plus pertinents.

5.4.2 Application multilingue sur quatre tâches

Dans un objectif de comparaison, nous avons appliqué cette méthodologie sur quatre tâches classiques du TAL (analyse de sentiments, reconnaissance d'entités nommées, reconnaissance d'implications textuelles (NLI), et reconnaissance de paraphrases) en deux langues (anglais et français) [Benamar et al., 2022a]. Le tableau 5.7 liste les quatorze corpus utilisés dans nos expériences, pour certains issus des référentiels FLUE et GLUE [20] (voir section 5.1.1).

Tâche	Corpus en anglais	Corpus en français
Analyse de sentiments	SST-2 (avis) IMDB (avis)	CLS-FR (avis produits) DEFT 2018 (tweets)
Reconnaissance d'entités nommées	Wiki-NER	Wiki-NER
Reconnaissance d'implications textuelles	MNLI-Matched MNLI-Mismatched	XNLI
Reconnaissances de paraphrases	Quora (QR) PIT (tweets)	DEFT 2020 (médical) OpusParcus PAWS-X

TABLE 5.7 : Corpus utilisés en anglais et en français sur les quatre tâches évaluées

Les modèles utilisés sont RoBERTa pour l'anglais et CamemBERT pour le français, dans leurs versions de base et large. Pour chacun de ces modèles, nous produisons une version enrichie des parties du discours, nommées « CamemBERT-POS » et « RoBERTa-POS ». Enfin, pour faciliter la comparaison des performances, nous avons effectué un affinage des huit modèles précédents.

Évaluation Le tableau 5.8 présente les résultats moyens obtenus en validation croisée 10-plis sur chaque tâche dans les deux langues, en fonction du modèle et de la version utilisés. Les résultats sont exprimés en F-mesure pour l'analyse de sentiments et la reconnaissance d'entités nommées, et par la corrélation de Pearson sur la reconnaissance de paraphrase et d'implications textuelles.

De manière globale, l'approche consistant à combiner des informations de parties du discours dans les représentations vectorielles des modèles RoBERTa et CamemBERT (version de base ou large) améliore les résultats sur trois des quatre tâches évaluées (reconnaissance de paraphrases, d'implications textuelles, et d'entités nommées), y compris par rapport à un affinage du modèle d'origine et de la version enrichie d'informations morpho-syntaxiques. En revanche, les performances baissent pour les deux langues sur la tâche d'analyse de sentiments, alors même que les corpus utilisés sont différents. Sur cette tâche, c'est la combinaison de l'affinage du modèle enrichi de parties du discours qui a permis d'améliorer les performances sur le corpus SST-2 en anglais.

20. CLS-FR, PAWS-X et XNLI sont issus de FLUE tandis que MNLI, Quora et SST-2 (Stanford Sentiment Treebank) proviennent de GLUE.

Modèle	Ana Sentiment		Reconnaissance Paraphrase			Implications		Entités
RoBERTa	SST-2	IMDB	Quora		PIT	MNLI	Mism.	Wiki
– base	0,57	0,87	0,70		0,36	0,43	0,47	0,85
– base(FT)	0,57	**0,89**	0,65		0,32	0,40	0,49	—
– base+POS	0,58	0,74	**0,76**		**0,38**	**0,45**	**0,53**	**0,87**
– base(FT)+POS	**0,59**	0,71	0,71		0,34	**0,45**	**0,53**	—
– large	0,54	0,88	0,72		**0,38**	0,52	0,54	0,87
– large(FT)	0,55	**0,89**	0,69		0,36	0,49	0,53	—
– large+POS	0,55	0,77	**0,74**		**0,38**	**0,55**	**0,56**	**0,91**
– large(FT)+POS	**0,56**	0,79	0,68		0,37	**0,56**	**0,56**	—
CamemBERT	CLS	DEFT-18	DEFT-20	OP	PAWS-X	XNLI		Wiki
– base	**0,86**	**0,54**	0,78	0,47	**0,70**	**0,74**		0,86
– base(FT)	0,81	0,52	0,81	0,45	0,63	0,69		—
– base+POS	0,74	0,47	**0,85**	0,53	0,67	**0,74**		**0,88**
– base(FT)+POS	0,73	0,42	0,83	**0,55**	0,68	0,72		—
– large	**0,86**	0,67	0,82	0,55	**0,71**	0,81		0,88
– large(FT)	0,84	**0,68**	0,72	0,47	0,66	0,79		—
– large+POS	0,69	0,65	**0,85**	**0,58**	**0,71**	**0,82**		**0,91**
– large(FT)+POS	0,70	0,66	0,80	0,47	0,69	**0,82**		—

TABLE 5.8 : Résultats obtenus en validation croisée 10-plis (F-mesure pour l'analyse de sentiments et la REN, corrélation de Person pour les paraphrases et implications textuelles) en fonction du modèle utilisé (RoBERTa sur l'anglais et CamemBERT sur le français, version base ou large, avec ajout de parties du discours, et/ou avec affinage (FT)). Les meilleurs résultats sont en gras

Conclusions Une première conclusion confirme l'intérêt d'enrichir les modèles existants avec des informations de parties du discours pour les tâches de reconnaissance de paraphrases, d'implications textuelles et d'entités nommées. En revanche, en raison de la dégradation observée des performances sur l'analyse de sentiments, ces informations ne sont pas suffisantes pour capturer la sémantique véhiculée par les opinions, sentiments et émotions (OSE), dans la mesure où les termes porteurs d'OSE se situent dans un cadre paradigmatique comme le soulignait Saussure [Saussure, 1916] (voir section 1.2). Une deuxième conclusion souligne l'intérêt de cette méthode d'enrichissement par rapport à un affinage, tant l'affinage des modèles disponibles que celui des modèles enrichis en parties du discours. À l'exception notable de l'analyse de sentiments (de nouveau), l'affinage des modèles n'a pas contribué à améliorer les performances de ces modèles.

5.5 Segmentation thématique de transcription de la parole

Je me suis récemment intéressé à la segmentation thématique automatique, dans la perspective de qualifier le contenu des productions langagières. J'ai réalisé ce travail sur un corpus de transcription automatique de la parole, dans la perspective finale d'analyser finement le contenu énoncé dans des émissions de radio et de télévision.

Pour identifier les différences de représentation femmes/hommes dans les médias et repérer les thématiques abordées, notamment celles concernées par une interruption de la parole, j'ai commencé à aborder la comparaison de plusieurs approches de segmentation thématique automatique appliquées sur des corpus de transcription automatique de la parole provenant d'émissions télévisées (voir page 20) pendant le stage de Lufei Liu. Deux difficultés font jour dans ce travail : la qualité contrastée de la transcription de la parole et l'impossibilité de transcrire de la parole superposée d'une part, et le manque de corpus de transcriptions annotés en thème d'autre part. Sur la base du corpus FrNewsLink [Camelin et al., 2018] annoté manuellement en thèmes et composé de 112 transcriptions de journaux télévisés français (du 10 au 16 février 2014 puis les 26 et 27

janvier 2015), nous avons adapté trois outils de segmentation thématique, représentatifs de plusieurs approches : TextTiling [Hearst, 1997] fondé sur l'identification de ruptures de distribution statistique de termes (une portion traite d'un nouveau thème si les mots employés diffèrent de ceux de la portion précédente), TopicTiling [Riedl and Biemann, 2012] permettant d'entraîner un modèle statistique, et DeepTiling [Ghinassi, 2021] fondé sur des réseaux de neurones.

Dans une première étape, nous avons identifié les frontières thématiques entre groupes de souffle. [21] Chaque bloc de trois groupes de souffle a fait l'objet d'une représentation vectorielle pour permettre une comparaison des performances des trois outils. L'évaluation a été mesurée en termes de couverture et de pureté, implémentées dans pyannote [Bredin, 2017], mesures qui tiennent compte des distances entre frontières de segment. Si TopicTiling obtient un score plus équilibré entre couverture (0,714) et pureté (0,822), c'est l'outil DeepTiling qui permet un découpage plus proche de la référence, avec une valeur très élevée de la couverture (0,922) mais la plus faible pureté (0,746) des trois outils testés.

La deuxième étape vise à identifier de quoi parle chaque portion thématique précédemment identifiée. Une preuve de concept a été réalisée, sur la base d'un entraînement effectué sur le corpus FrNewsLink de 2014, pour application sur le corpus FrNewsLink de 2015. Les termes mis en avant par l'approche utilisée ne permettent pas de qualifier le contenu de chaque portion thématique. Parmi les erreurs relevées, nous observons que des termes extraits correspondent à des passages publicitaires, ou à la fin d'une émission précédente. Un autre constat révèle que les annotateurs humains choisissent des termes différents de ceux employés dans le groupe de souffle considéré pour nommer la thématique, ce qui rend complexe une évaluation automatique. Ce travail, à peine entamé, mériterait d'être poursuivi, tant il révèle des problématiques nombreuses.

5.6 Conclusion

Ce chapitre concentre les recherches que j'ai effectuées autour des modèles pré-entraînés fondés sur l'architecture des transformers, principalement pendant la thèse d'Alexandra Benamar. Ces travaux ont porté sur les productions langagières composées de mots hors vocabulaire, par définition absents des vocabulaires des modèles transformers. Nous pouvons distinguer deux sources de production de ces mots HV.

Une première origine concerne à la fois les erreurs orthographiques et fautes de frappe produites par les utilisateurs dans la communication numérique, mais également les termes qui sont spécifiques à un type de communication (les échanges électroniques formels dans le corpus de mails EDF) ou à un domaine de spécialité ou une thématique donnée. Nous avons mis en évidence des différences de représentation du texte en fonction du corpus qui a servi à l'entraînement des modèles. Ainsi, la tokénisation en sous-unités effectuée par le modèle CamemBERT est plus pertinente sur le corpus de mails EDF si le modèle a été entraîné sur les corpus CCNet et OSCAR (composés de pages web) que sur des pages Wikipédia (section 5.2). Nous avons également vérifié que les modèles n'encodent pas que la sémantique contenue dans les corpus d'entraînement, mais également des « biais » que nous avons étudiés sous l'angle des stéréotypes de genre, sur un corpus de pièces de théâtre classique parmi deux types de pièces (comédie et tragédie). Comme pour la tokénisation en sous-unités, nous avons vérifié par l'analyse des voisins distributionnels que le corpus utilisé en entraînement véhicule des stéréotypes de genre différents et des émotions genrées (section 5.3.2).

Enfin, et parce que ces modèles pré-entraînés sont coûteux à utiliser, nous avons étudié une solution alternative au réentraînement et à l'affinage aux données, en nous fondant sur une retokénisation des sous-unités générées par les modèles en combinant des parties du discours (section 5.4). Nous avons evalué l'intérêt de cette démarche sur l'anglais et le français, pour quatre tâches classiques du traitement automatique des langues. Nous avons constaté que cette approche

21. Parce qu'il n'existe pas de ponctuation en parole permettant d'identifier des phrases terminées par un point, notre unité de travail est le groupe de souffle, défini comme une séquence parlée entre deux silences.

permet une amélioration des performances sur trois tâches (reconnaissance de paraphrases, d'implications textuelles et d'entités nommées) alors qu'elle peut contribuer à dégrader les résultats en analyse de sentiments, tant sur l'anglais que sur le français. Je suppose que la limite de l'approche testée est liée au fait que les termes porteurs d'opinion, sentiment, et émotion relèvent des mêmes classes de parties du discours (donc cette information n'est pas discriminante pour cette tâche) et qu'ils évoluent dans une relation paradigmatique, avec un antagonisme qui ne peut pas être capturé, ou pas uniquement, par une information morpho-syntaxique. Sur cette tâche, la combinaison d'information de valence pourrait constituer une solution alternative à l'affinage, en reprenant la méthode développée par Alexandra Benamar pour l'ajout des parties du discours. Des listes de termes avec information de valence comme TreeLex pour les verbes et adjectifs [Kupść, 2008, Kupść and Abeillé, 2008] et des approches permettant de générer de tels dictionnaires pour le français ont été proposées [Vincze and Bestgen, 2011].

Chapitre 6

Conclusion et perspectives

Sommaire

6.1 Conclusion

Dans ce manuscrit, je suis revenu sur les recherches que j'ai menées ces dernières années autour des productions langagières des utilisateurs des réseaux sociaux, en écartant les recherches que j'ai pu effectuer à partir de textes normés et produits par des spécialistes d'un domaine (tels que les comptes-rendus hospitaliers qui correspondent à une structure fixe). J'ai analysé ce type de production sous deux angles différents.

6.1.1 Des choix individuels qui transforment la langue

Dans une première partie, j'ai considéré ces productions comme un objet d'étude à interroger (chapitre 2), du point de vue des éléments linguistiques et méta-linguistiques identifiables, en mettant en avant les hétérogénéités liées à la langue (section 2.2) de celles liées aux usages spécifiques aux réseaux sociaux (section 2.3). Alors que certains aspects linguistiques s'imposent aux locuteurs d'une langue (système d'écriture, translittération des mots étrangers), d'autres relèvent d'un choix délibéré desdits locuteurs, qui peuvent être liés à des possibilités techniques (code switching, ajout d'émojis et de hashtags) ou à une volonté politique de l'émetteur du message (utilisation militante du français inclusif). Les travaux de thèse de Liyun Yan consacrés à l'analyse des messages laissés par les touristes chinois en visite à Paris, ou les expériences sur les corpus de transcription automatique de la parole sur la télé-réalité, permettent de prendre conscience de la diversité des spécificités propres à chaque corpus, et de l'apparente difficulté qui émerge des différentes productions.

Il importe cependant de relativiser ces observations, dans la mesure où l'argot, le *verlan*, et des sous-langages ont toujours existé, notamment au sein d'une communauté pour coder le contenu des échanges vis à vis des personnes n'appartenant pas à cette communauté (exemple typique du *loucherbem* utilisé par les bouchers de La Villette au XIX$^{\text{ème}}$ siècle et dont certains termes sont passés dans le langage argotique courant et continuent d'être utilisés un siècle et demi plus tard : « *en loucedé* »). Ces choix sont restés limités à une communauté et ne semblent pas avoir modifié durablement la langue. Les changements liés aux évolutions techniques peuvent cependant avoir un impact plus durable, non pas dans la langue, mais au niveau des usages. Je pense notamment à l'intégration d'émojis et de hashtags, qui se fait dans la chaîne écrite, mais ne se traduit pas par un remplacement dans le vocabulaire, et de manière plus limitée pour la chaîne parlée, en tant qu'expression permettant de ponctuer oralement des anecdotes : « *hashtag malaise* » ou « *hashtag JPP* » (« *j'en peux plus* ») pour expliciter une émotion.

Alors que des différences culturelles existent dans l'expression des opinions, j'ai ensuite présenté les travaux que nous avons effectués avec Liyun Yan autour de la fouille d'opinion sur ce corpus d'avis de touristes chinois. J'ai axé ce travail autour des non-dits des utilisateurs (chapitre 3) en travaillant autour des inférences. Dans un premier temps, je me suis intéressé avec Liyun à

proposer une classification des inférences (section 3.2) autour de trois dimensions (réalisation sémantique, modalité de réalisation, et mode de production), travail qui s'inspire de celui de Peirce réalisé sur les déduction, induction et abduction. Nous avons toutefois constaté la complexité que cela représente pour une annotation manuelle, et fait le choix d'abandonner la modalité de réalisation dans les expériences de fouille d'opinion réalisées sur ce corpus (section 3.3). Les expériences réalisées à base d'apprentissage statistique, fondées sur les informations sémantiques précédemment annotées, nous ont permis d'identifier aussi bien la présence d'inférence dans un message que le type de réalisation sémantique (logique, pragmatique, ou lexicale) que le mode de production (énonciatif ou discursif). Nous avons par ailleurs observé que les inférences constituent, pour la fouille d'opinion, un moyen strictement complémentaire à l'identification des mots porteurs d'opinion/sentiment/émotion. De manière plutôt inattendue, j'ai également constaté que les performances de modèles appris sur des sinogrammes simplifiés ne dénotent pas franchement de celles obtenus sur un corpus rédigé en alphabet latin (pinyin), alors que le nombre de caractères est sans commune mesure. Devant la perte d'information sémantique liée au passage en pinyin, nous avons conclu qu'il était préférable de conserver les sinogrammes, d'autant plus que les systèmes actuels du TAL peuvent les traiter.

6.1.2 Des évolutions qui se retrouvent dans les ressources employées en TAL

Dans une deuxième partie, je me suis intéressé à l'impact que les productions langagières des utilisateurs peut constituer sur les ressources créées pour le TAL, et sur les performances des outils du TAL entraînés sur ces données. Si la présence des évolutions langagières dans les ressources employées en TAL ne constitue pas une surprise, elle interroge cependant la communauté sur la manière de prendre en compte ces évolutions.

J'ai consacré un chapitre de ce manuscrit à l'utilisation des réseaux sociaux dans un objectif de pharmacovigilance (chapitre 4), dans lequel j'ai pu constater que les propriétés idiosyncrasiques des corpus observées dans les précédents chapitres se retrouvent dans ce domaine d'application. Dans un premier temps, j'ai focalisé mes recherches sur la comparaison des modélisations possibles des informations médicales, en analysant plusieurs schémas d'annotation (section 4.2), tant du point de vue de l'annotation humaine (d'un schéma global à un schéma nucléaire avec des relations d'expansion, en étudiant la possibilité de passer de l'un à l'autre) que des performances obtenues par des systèmes de TAL. Au-delà des aspects modélisations, je me suis également intéressé avec François Morlane-Hondère aux indices linguistiques qui permettent de faciliter le repérage des informations utiles pour la pharmacovigilance (section 4.3, verbes introducteurs de médicaments, typage des relations causales, utilisation des voisins distributionnels, etc.). Alors même que le repérage d'informations spécialisées se révèle complexe sur des textes normés, j'ai pu mettre en évidence que le même repérage sur la production des locuteurs bénéficie de la combinaison des informations linguistiques avec les approches par apprentissage statistique. Cette observation n'a cependant pas pu être vérifiée sur la recherche de témoignage de mésusage médicamenteux (section 4.4), en raison de la complexité à identifier une consommation déviante de médicament en termes de détournement d'usage ou de posologie. Le lien avec des connaissances médicales, réalisé par les médecins, suppose une compréhension fine des contenus renseignés sur les réseaux sociaux, sans qu'il ne soit possible de revenir vers la personne à l'origine du témoignage pour obtenir des informations complémentaires.

Dans mon dernier chapitre, je reviens sur l'étude des modèles à base de plongements (chapitre 5), tant du point de vue des informations véhiculées par ces modèles que des possibilités d'amélioration pour éviter un réentraînement ou un affinage potentiellement coûteux. Sur la thèse d'Alexandra Benamar, j'ai observé que la tokénisation en sous-unités imposée par ces modèles représente différemment les mots hors vocabulaire en fonction du corpus qui a servi à entraîner ces modèles (section 5.2). Nous avons également confirmé la présence de stéréotypes de genre en affinant des modèles sur un corpus de pièces du théâtre classique parmi deux genres (comédie et

tragédie), en partant du principe que ces genres exagèrent certains traits, dont les stéréotypes de genre (section 5.3). Parce que les questions environnementales émergent autour de la création et de l'utilisation de modèles transformers, et que des propositions de verdissement de ces modèles font jour, j'ai orienté Alexandra sur le moyen d'éviter un réentraînement ou un affinage coûteux (section 5.4). Grâce au travail d'ajout d'informations de parties du discours après retokénisation des sous-unités générées, j'ai pu étudié l'intérêt de cette démarche sur quatre tâches classiques du TAL. Si les résultats sont améliorés par cette méthode sur les tâches de reconnaissance d'entités nommées, d'implications textuelles, et de paraphrases, nous avons constaté que l'analyse de sentiments ne bénéficie pas de cet apport d'informations morpho-syntaxiques. J'attribue cette absence de résultat au fait que les termes porteurs d'opinion/sentiment/émotion se situent dans un cadre purement paradigmatique. Alors que l'impact des locuteurs d'une langue apparaît clairement sur les réseaux sociaux lors de l'application d'approches à base de TAL, la révolution apportée par les plongements et les transformers s'accompagne d'une prise de conscience du fait que ces modèles n'encodent pas que la sémantique contenue dans les corpus utilisés pour entraîner ces modèles, mais que tous les aspects sociétaux véhiculés par la langue se trouvent également encodés. Il y a quelques années, la présence de ces éléments sociétaux aurait constitué un domaine d'étude. Aujourd'hui, cette présence interroge la communauté de l'IA autour de questions éthiques relatives à l'utilisation de ces modèles, comme pour l'éphémère chatbot @Tay sur les réseaux sociaux.

6.2 Perspectives

Alors que j'ouvrais ce manuscrit en témoignant d'un constat effectué autour des problèmes qualifiés de « résolus » en traitement automatique des langues, les travaux que j'ai présentés, tant autour de l'objet d'étude en lui-même, que de l'utilisation des productions langagières pour produire de nouvelles ressources et modèles, vont dans le sens contraire (y compris sur des tâches de base comme une tokénisation lors du traitement du mandarin standard), et rappellent que la langue est un objet vivant qui évolue constamment au grés des évolutions technologiques (nécessitant de nommer de nouvelles réalités) et sociétales (notamment en lien avec l'usage, essentiellement à l'oral, du français inclusif), et que les modèles statistiques actuels (voire futurs) ne pourront jamais qu'approximer l'encodage de la sémantique globale conceptualisée par Harris et Chomsky. Est-il cependant possible d'en être sûr ?

6.2.1 Les cycles de la recherche : suivre les évolutions langagières

Il serait plus juste de voir et concevoir les recherches en traitement automatique des langues comme s'inscrivant dans des cycles, à l'image des études qui existent en sciences sociales et sciences économiques, liées aux cycles sociétaux ou économiques [Catroux, 2002, Wood, 2008], ou plus généralement dans la démarche scientifique [Beaugrand, 1988] [1] et formalisée pour mettre en place des méthodologies expérimentales [Anceaux, 2006]. Chaque début de cycle correspond à une évolution majeure dans la langue ou dans les propriétés des données à traiter—je reprends ici l'intégration des émojis et hashtags dans la chaîne écrite ou encore la démocratisation du français inclusif, mais également la pandémie de Covid-19 qui aura contribué à focaliser les discussions sur cette thématique sur un temps long, venant perturber la distribution des termes du vocabulaire par rapport à la période antérieure à cette pandémie—, et se termine par la résolution des problèmes posés par cette évolution, jusqu'au démarrage d'un nouveau cycle porté par une nouvelle évolution technique ou sociétale qui aura un impact sur la langue. Dans cette perspective, on peut considérer l'arrivée des modèles de type BERT comme point de départ d'un nouveau cycle, contribuant à dynamiser les recherches actuelles en TAL et en IA autour de ce type de modèle.

1. Beaugrand distingue quatre étapes dans un cycle : (*i*) la préparation de la production des observations et mesures, (*ii*) la production des observations, (*iii*) l'analyse et l'interprétation des données conduisant éventuellement à questionner et reformuler le modèle, et (*iv*) la publication.

Lors de la rédaction de ce chapitre de perspectives en septembre 2022, je m'interrogeais sur la révolution attendue de l'informatique quantique qui devrait, selon toutes vraisemblances, remettre en cause ma certitude sur l'impossiblité d'encoder la sémantique globale d'une langue, en raison du volume toujours plus élevé de calculs à réaliser sur un volume de données en expansion constante. J'envisageais cette possibilité pour traiter des contextes de plus en plus importants, jusqu'à pouvoir traiter l'intégralité d'un document, voire un ensemble de documents, ce qui se révélerait utile pour la compréhension ou la prédiction d'effets, notamment en domaine médical, pour croiser des informations réparties entre plusieurs documents d'un même dossier patient.

Entre-temps, le modèle ChatGPT [2] a été rendu public en novembre 2022 par la société OpenAI. Je reconnais que les performances de ce prototype de chatbot sont, de prime abord, réellement impressionnantes, notamment pour ce qui concerne l'analyse linguistique effectuée (y compris la prise en compte des reprises anaphoriques dans le dialogue), et ont donné lieu à plusieurs discussions enthousiastes à la cafétéria du LISN. Ces performances ont également été remarquées par les enseignants qui craignent que les étudiants se servent de cette intelligence artificielle pour produire du contenu. En réponse, OpenAI annonce en décembre 2022 travailler sur l'ajout d'une signature numérique *(digital watermark)* sur les contenus produits par son modèle pour les identifier plus clairement. Mais soyez rassurés, je certifie que l'intégralité de ce manuscrit a bien été produit par un humain.

L'arrivée de ce modèle semble acter un changement de paradigme dans la formalisation des connaissances entre humains et machines. Si les précédentes recherches visaient à intégrer du TAL dans les systèmes informatiques pour une tâche donnée, en se fondant sur la représentation humaine des connaissances, les capacités informatiques actuelles permettent désormais de s'abstraire de ces représentations linguistiques pour aboutir aux mêmes résultats qu'un humain. Pour autant, les recherches en TAL ne sont pas totalement dépassées. Même si ce modèle ChatGPT ne résoud pas tous les problèmes, il devra, lui aussi, faire l'objet d'adaptations aux nouveaux usages des locuteurs et aux évolutions de la langue comme je l'indiquais en introduction générale pour les approches plus classiques.

6.2.2 Poursuivre l'intégration d'informations linguistiques dans les plongements

Malgré tout, une interrogation demeure quant à la possibilité de capturer la diversité des propriétés linguistiques présentes dans les productions langagières. Une question de recherche que j'ai commencé à aborder avec Alexandra Benamar concerne le lien entre les modèles statistiques utilisés en TAL et les informations linguistiques issues des données. Si les observations statistiques permettent de capturer des régularités, qui peuvent se traduire dans la langue, je milite pour que le TAL ne se réduise pas à la seule statistique de phénomènes observés, et que l'intégration d'informations fines au niveau linguistique se poursuive, même s'il s'agit d'un travail laborieux qui ne se traduit pas rapidement par des effets de masse, comme c'est le cas pour les modèles transformers actuels. Pour le dire ouvertement, je ne souhaite pas que le TAL soit fongible dans la science des données.

En lien avec la thématique de la désidentification que je n'ai pas abordée dans ce manuscrit, mais qui constitue pour moi un champ de recherche toujours actif, alors que nous avons pu mettre en évidence la présence de stéréotypes de genre dans les représentations vectorielles, la présence d'informations nominatives et potentiellement identifiantes dans des modèles qui pourraient être entraînés sur des corpus de documents cliniques doit être envisagée sérieusement. En effet, les inquiétudes soulevées par la communauté sur le mauvais emploi des modèles transformers pour générer des propos misogynes ou racistes provenant des informations encodées dans lesdits modèles (« biais ») et qui relèvent d'une potentialité, ne doivent pas occulter un problème à mes yeux plus sérieux qui reste celui de la confidentialité des informations présentes dans ces modèles, et dont la

2. https://openai.com/blog/chatgpt/

présence me semble attestée. J'ai commencé à aborder cette problématique il y a quelques années, en me focalisant sur les approches délexicalisées pour produire des modèles CRF redistribuables. Sans aller sur le darkweb, il est possible de trouver sur internet des documents cliniques réels, déposés directement par les patients, ou provenant d'erreur de manipulation, et pour desquels les approches statistiques actuelles ou à venir pourraient encoder des informations sensibles, avec un préjudice certain pour les personnes concernées (piratage, escroquerie, etc.).

Si l'ajout d'information linguistique dans les plongements utilisés par les modèles *transformers* a été expérimenté pendant la thèse d'Alexandra Benamar avec les parties-du-discours, et réalisé par d'autres équipes avec une combinaison d'informations (étiquettes en partie du discours, information de casse typographique, position de la sous-unité, etc.) [Sundararaman et al., 2021] [3] sur le référentiel GLUE, conduisant à une amélioration des résultats, je serai intéressé de combiner ces informations dans mes activités de désidentification de textes cliniques. J'envisage notamment de voir ce que pourrait donner l'intégration d'informations de valence et d'opinion-sentiment-émotion dans les plongements. Si le transfert de connaissance a été abordé en cas d'absence d'annotations sur plusieurs tâches et pour plusieurs langues [Lauscher et al., 2020], soit en provenance d'un autre domaine (du légal vers le clinique), soit d'une autre langue relativement proche (de l'anglais au français), je souhaiterai vérifier l'opportunité d'y avoir recours pour compléter les informations disponibles dans les représentations utilisées par les modèles *transformers*.

6.2.3 Ou revenir vers plus de simplicité

Par opposition aux démarches actuelles reposant sur les plongements et l'utilisation de modèles *transformers* dont on a souligné l'importance des coûts énergétiques, environnementaux et financiers dans les différentes étapes d'utilisation [Poerner et al., 2020], revenir vers des solutions plus simples me semblent encore pertinent. Toujours sur la thématique de la désidentification, mais que l'on peut extrapoler comme un repérage d'entités nommées, je souhaiterai étudier et vérifier l'opportunité, non pas de repérer les éléments identifiants dans un texte (objet de la désidentification), mais d'identifier les portions de texte (quelle que soit la granularité, de la portion de phrase au paragraphe complet) qui ne contiennent aucune données identifiantes. Il s'agit là d'une vision complémentaire à la désidentification et qui consiste à assurer que les portions repérées ne contiennent aucune information identifiante.

Alors que le TAL biomédical a besoin de données textuelles cliniques pour mettre au point des modèles et outils adaptés aux spécificités de cette langue de spécialité, j'estime que la désidentification ne constitue qu'une solution possible pour mettre à disposition du matériau textuel représentatif du domaine [4]. J'estime que l'identification de portions textuelles ne contenant aucune information identifiante constitue une solution valable dans la mesure où ce sont précisément ces portions qui proposent des exemples d'usages langagiers propres à la langue biomédicale [Zweigenbaum et al., 2001], et non les passages fournissant des informations démographiques (nom, adresse, date de naissance, etc.). Un autre objectif du repérage de ces portions concerne la possibilité de fournir ces formulations produites par les professionnels du soin comme modèle, soit pour générer des textes du domaine (sur lesquels entraîner des outils du TAL biomédical), soit pour fournir une base de rédaction à ces mêmes professionnels en vue d'une réduction du temps de production des textes cliniques. Cette dimension constitue une recherche que j'envisage de mener à court terme avec une orthophoniste amenée à rédiger régulièrement des comptes-rendus et qui s'est déjà interrogée sur sa pratique langagière [Brin-Henry, 2014], mettant en avant un besoin d'homogénéisation de la terminologie orthophonique utilisée dans les productions scientifiques [Brin-Henry, 2019], et par extension dans les textes cliniques.

3. Les auteurs ont également mené une réflexion sur la manière d'encoder ces informations dans les plongements. Le mot anglais « *amalgamation* » est représenté par trois sous-unités : 'amal,NOUN,0,S', 'ga,NOUN,0,M', et 'mation,NOUN,0,E', avec 0 pour la casse en minuscules et S/M/E la position de la sous-unité (start/middle/end).

4. Une exception notable à cette assertion concerne précisément le besoin d'exemples réels pour mettre au point des systèmes de désidentification.

Je souhaiterai néanmoins terminer ce chapitre de perspectives et ce manuscrit sur une note plus positive. Parce que la langue reste un objet social, et parce que les sociétés évoluent continuellement, nous pouvons être assurés que les chercheurs en traitement automatique des langues auront toujours des problématiques de recherche à aborder, et des problèmes qu'ils pourront considérer comme résolus à un moment donné, jusqu'à ce qu'une nouvelle génération prenne la relève et soulève de nouveaux problèmes.

[Alpheratz, 2018] Alpheratz (2018). Français inclusif : conceptualisation et analyse linguistique. In *SHS Web Conf, Congrès Mondial de Linguistique Française*, volume 46. doi:10.1051/shsconf/20184613003. – Cité page 27.

[Alpheratz, 2019] Alpheratz (2019). Français inclusif : du discours à la langue ? *Le discours et la langue*, 1(111). hal-02323626/. – Cité pages 27 et 31.

[Anceaux, 2006] Anceaux, Françoise, e. P. S. (2006). Mise en place d'une méthodologie expérimentale : hypothèses et variables. *Recherche en soins infirmiers*, 84(1):66–83. doi:10.3917/rsi.084.0066. – Cité page 83.

[Araujo and Kollat, 2018] Araujo, T. and Kollat, J. (2018). Communicating effectively about CSR on Twitter: the power of engaging strategies and storytelling elements. *Internet Research*, 28(2):419–431. doi:10.1108/intr-04-2017-0172. – Cité page 15.

[Arrivé et al., 1986] Arrivé, M., Gadet, F., and Galmiche, M. (1986). *La grammaire d'aujourd'hui*. Flammarion, Paris. – Cité page 28.

[Audeh et al., 2019] Audeh, B., Grouin, C., Zweigenbaum, P., Bousquet, C., Jaulent, M.-C., Benkhebil, M., and Lillo-Le Louët, A. (2019). French Levothyrox® Crisis: retrospective analysis of social media. In *International Society of Pharmacovigilance*, Bogota, Colombia. Springer International Publishing. hal-02411632. – Cité page 51.

[Austin, 1962] Austin, J. L. (1962). *How to Do Things with Words*. Clarendon Press, Oxford. The William James Lectures delivered in Harvard University in 1955. – Cité page 27.

[Azé et al., 2006] Azé, J., Heitz, T., Mela, A., Mezaour, A.-D., Peinl, P., and Roche, M. (2006). Présentation de DEFT'06 (DEfi Fouille de Textes). In *Actes de DEFT*, Fribourg, Suisse. lirmm-00113164v2. – Cité page 69.

[Baghdadi et al., 2019a] Baghdadi, Y., Bourrée, A., Robert, A., Rey, G., Gallay, A., Zweigenbaum, P., Grouin, C., and Fouillet, A. (2019a). Automatic classification of free-text medical causes from death certificates for reactive mortality surveillance in France. *International Journal of Medical Informatics*, 131:103915. doi:10.1016/j.ijmedinf.2019.06.022. – Cité page 10.

[Baghdadi et al., 2019b] Baghdadi, Y., Bourrée, A., Robert, A., Rey, G., Gallay, A., Zweigenbaum, P., Grouin, C., and Fouillet, A. (2019b). A new approach to compare performance of two classification methods of causes of death for timely surveillance in France. *Studies in Health Technology and Informatics*, 264:925–929. doi:10.3233/shti190359. – Cité page 10.

[Baghdadi et al., 2019c] Baghdadi, Y., Bourrée, A., Robert, A., Rey, G., Gallay, A., Zweigenbaum, P., Grouin, C., and Fouillet, A. (2019c). Performance of machine-learning method to classify free-text medical causes of death. *Online Journal of Public Health Informatics*, 11(1):e258. doi:10.5210/ojphi.v11i1.9767. – Cité page 10.

[Barre et al., 2020] Barre, A., Bordage, P., Faye, E., Chevalier, C., Ecken, C., Granier de Cassagnac, R., Berry, A., Lainé, S., Debats, J.-A., Ligny, J.-M., Genefort, L., Bihouix, P., Viguié, V., Moutou, F., Moreira, V., Husset, M.-J., Auzanneau, M., Chenu, C., Pellerin, S., Czernichowski-Lauriol, I., Maugis, P., Lecomte, J., Dufour, C., and Sarrazi, F. (2020). *Nos futurs. Imaginer les Possibles du Changement Climatique*. Éditions ActuSF. https://www.editions-actusf.fr/a/collectifd-auteur/nos-futurs. – Cité page 31.

[Beaugrand, 1988] Beaugrand, J. P. (1988). Démarche scientifique et cycle de la recherche. In Robert, M., editor, *Fondements et étapes de la recherche scientifique en psychologie*, pages 1–35. Edisem-Maloine, Saint-Hyacinthe, Canada. – Cité page 83.

[Belainine et al., 2016] Belainine, B., Fonseca, A., and Sadat, F. (2016). Named entity recognition and hashtag decomposition to improve the classification of tweets. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 102–111, Osaka, Japan. The COLING 2016 Organizing Committee. https://aclanthology.org/W16-3915. – Cité page 26.

[Beltagy et al., 2019] Beltagy, I., Lo, K., and Cohan, A. (2019). SciBERT: a pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics. doi:10.18653/v1/D19-1371. – Cité page 66.

[Ben-Lhachemi and Nfaoui, 2017] Ben-Lhachemi, N. and Nfaoui, E. (2017). An extended spreading activation technique for hashtag recommendation in microblogging platforms. In *Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics*, pages 1–8, Amantea, Italy. Association for Computing Machinery. doi:10.1145/3102254.3102283. – Cité page 26.

[Benamar et al., 2021] Benamar, A., Bothua, M., Grouin, C., and Vilnat, A. (2021). Easy-to-use combination of POS and BERT model for domain-specific and misspelled terms. In *NL4IA Workshop Proceedings*, Milan, Italy. hal-03474696. – Cité pages 10, 71 et 76.

[Benamar et al., 2022a] Benamar, A., Grouin, C., Bothua, M., and Vilnat, A. (2022a). Adding morpho-syntax: a simple yet effective approach to adapt word representation for small domain-specific datasets using transformers. In *Soumission ARR en cours*. – Cité page 77.

[Benamar et al., 2022b] Benamar, A., Grouin, C., Bothua, M., and Vilnat, A. (2022b). Étude des stéréotypes genrés dans le théâtre français du XVI^{e} au XIX^{e} siècle à travers des plongements lexicaux. In Estève, Y., Jiménez, T., Parcollet, T., and Zanon Boito, M., editors, *Traitement Automatique des Langues Naturelles*, pages 74–81, Avignon, France. ATALA. hal-03701478. – Cité pages 10 et 73.

[Benamar et al., 2022c] Benamar, A., Grouin, C., Bothua, M., and Vilnat, A. (2022c). Evaluating tokenizers impact on OOVs representation with transformers models. In *Proceedings of the Language Resources and Evaluation Conference*, pages 4193–4204, Marseille, France. European Language Resources Association. https://aclanthology.org/2022.lrec-1.445. – Cité pages 10 et 72.

[Benamara et al., 2017] Benamara, F., Grouin, C., Karoui, J., Moriceau, V., and Robba, I. (2017). Analyse d'opinion et langage figuratif dans des tweets : présentation et résultats du défi fouille de textes DEFT2017. In *Atelier TALN 2017 : Défi Fouille de Textes (DEFT 2017)*, pages 1–12, Orléans, France. Association pour le Traitement Automatique des Langues. hal-01912785. – Cité pages 16 et 25.

[Benveniste, 1976] Benveniste, E. (1976). De la subjectivité dans le langage. In *Problèmes de linguistique générale, 1*, chapter XXI, pages 258–266. Gallimard, Paris. Parution initiale dans le *Journal de Psychologie*, juil.-sept. 1958, P.U.F. – Cité pages 8 et 34.

[Benveniste, 1980] Benveniste, E. (1980). L'appareil formel de l'énonciation. In *Problèmes de linguistique générale, 2*, chapter V, pages 79–88. Gallimard, Paris. Parution initiale dans *Langages*, n° 17, mars 1970, Didier-Larousse. – Cité page 33.

[Berardi et al., 2011] Berardi, G., Esuli, A., Marcheggiani, D., and Sebastiani, F. (2011). ISTI@TREC Microblog track 2011: exploring the use of hashtag segmentation and text quality ranking. In *Proceedings of The Twentieth Text REtrieval Conference (TREC 2011)*, Gaithersburg, MD. https://trec.nist.gov/pubs/trec20/papers/NEMIS_ISTI_CNR.microblog.update.pdf. – Cité page 26.

[Bigeard et al., 2018] Bigeard, E., Grabar, N., and Thiessard, F. (2018). Detection and analysis of drug misuses. a study based on social media messages. *Frontiers in Pharmacology*, 9. doi:10.3389/fphar.2018.00791. – Cité pages 62 et 63.

[Biscarrat et al., 2022] Biscarrat, L., Doukhan, D., and Grouin, C. (2022). 20 ans de télé-réalité, 20 ans de sexisme ? De Loft Story aux Marseillais à Dubaï : apport des méthodes d'analyse automatique pour la description des évolutions du dispositif télévisuel. In *90ème Congrès de l'ACFAS*, Montréal, Canada. – Cité page 20.

[Blodgett et al., 2020] Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. (2020). Language (technology) is power: a critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online symposium. Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.485. – Cité page 73.

[Bœuf et al., 2017] Bœuf, M., Bellet, F., Karapetiantz, P., Leprovost, D., Morlane-Hondère, F., Grouin, C., Audeh, B., Bousquet, C., and Beyens, M.-N. (2017). A pilot study of the Vigi4MED project: comparison of adverse drug reactions (ADRs) of duloxetine between patients' forum posts and the French pharmacovigilance database (FPVD). *Fundamental & Clinical Pharmacology*, 31:33. – Cité page 51.

[Bojanowski et al., 2017] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146. doi:10.1162/tacl_a_00051. – Cité page 65.

[Bredin, 2017] Bredin, H. (2017). pyannote.metrics: a toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden. https://-pyannote.github.io/pyannote-metrics. – Cité page 79.

[Bridge et al., 2021] Bridge, O., Raper, R., Strong, N., and Nugent, S. E. (2021). Modelling a socialised chatbot using trust development in children: lessons learnt from Tay. *Cognitive Computation and Systems*, 3(2):100–108. doi:10.1049/ccs2.12019. – Cité page 73.

[Brin-Henry, 2014] Brin-Henry, F. (2014). Using corpus-based analyses in specialised paramedical French. *Revue française de linguistique appliquée*, XIX(1):103–115. doi:10.3917/rfla.191.0103. – Cité page 85.

[Brin-Henry, 2019] Brin-Henry, F. (2019). Pour une harmonisation de la terminologie orthophonique : contribution du projet OrthoCorpus (2015–2017). Rapport de projet, Analyse et Traitement Informatique de la Langue Française (ATILF). hal-02330551. – Cité page 85.

[Brun and Roux, 2014] Brun, C. and Roux, C. (2014). Décomposition des « hash tags » pour l'amélioration de la classification en polarité des « tweets ». In *Proceedings of TALN 2014 (Volume 2 : Short Papers)*, pages 473–478, Marseille, France. Association pour le Traitement Automatique des Langues. https://aclanthology.org/F14-2015. – Cité page 26.

[Camelin et al., 2018] Camelin, N., Damnati, G., Bouchekif, A., Landeau, A., Charlet, D., and Estève, Y. (2018). FrNewsLink : a corpus linking TV broadcast news segments and press articles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA). https://aclanthology.org/L18-1329. – Cité page 78.

[Campillos-Llanos et al., 2019] Campillos-Llanos, L., Grouin, C., Lillo-Le Louët, A., and Zweigenbaum, P. (2019). Initial experiments for pharmacovigilance analysis in social media using summaries of product characteristics. *Studies in Health Technology and Informatics*, 264:60–64. doi:10.3233/shti190183. – Cité pages 10 et 61.

[Cardon et al., 2020] Cardon, R., Grabar, N., Grouin, C., and Hamon, T. (2020). Présentation de la campagne d'évaluation DEFT 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques. In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement*

Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes, Nancy, France. hal-02784737. – Cité page 50.

[Carlotti, 2011] Carlotti, A. (2011). *Phrase, énoncé, texte, discours. De la linguistique universitaire à la grammaire scolaire*. Lambert-Lucas, Limoges. – Cité page 33.

[Catroux, 2002] Catroux, M. (2002). Introduction à la recherche-action : modalités d'une démarche théorique centrée sur la pratique. *Recherche et pratiques pédagogiques en langues de spécialité — Cahiers de l'APLIUT*, XXI(3):8–20. doi:10.4000/apliut.4276. – Cité page 83.

[Cattan et al., 2021] Cattan, O., Servan, C., and Rosset, S. (2021). On the usability of transformers-based models for a French question-answering task. In *Recent Advances in Natural Language Processing, RANLP 2021*, Online. hal-03336060. – Cité page 67.

[Ceusters et al., 1998] Ceusters, W., Buekens, F., De Moor, G., and Waagmeester, A. (1998). The distinction between linguistic and conceptual semantics in medical terminology and its implication for NLP-based knowledge acquisition. *Methods for Informatics in Medicine*, 37(4–5). PMID:9865030. – Cité page 63.

[Chalkidis et al., 2020] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020). LEGAL-BERT: the Muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics. doi:10.18653/v1/2020.findings-emnlp.261. – Cité page 66.

[Chang and Lin, 2011] Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol*, 2(3):1–27. doi:10.1145/1961189.1961199. – Cité page 57.

[Chapman et al., 2001] Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., and Buchanan, B. G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34(5). doi:10.1006/jbin.2001.1029. – Cité page 23.

[Charaudeau, 1984] Charaudeau, P. (1984). Une théorie des sujets du langage. *Langage et société*, 28(1):37–51. doi:10.3406/lsoc.1984.1989. – Cité page 33.

[Chen et al., 2018] Chen, Y., Yuan, J., You, Q., and Luo, J. (2018). Twitter sentiment analysis via bi-sense emoji embedding and attention-based LSTM. In *Proceedings of the 26th ACM international conference on Multimedia*, Seoul, Corée du Sud. Association for Computing Machinery. doi:10.1145/3240508.3240533. – Cité page 25.

[Chiasson, 2005] Chiasson, P. (2005). Abduction as an aspect of retroduction. *Semiotica*, 2005(153):223–242. doi:10.1515/semi.2005.2005.153-1-4.223. – Cité page 35.

[Chomsky, 1957] Chomsky, N. (1957). *Syntactic Structure*, volume 4. De Gruyter Mouton. doi:10.1515/9783112316009. – Cité pages 8 et 33.

[Chong Ho, 1994] Chong Ho, Y. (1994). Abduction? Deduction? Induction? Is there a Logic of Exploratory Data Analysis? In *Annual Meeting of American Educational Research Association*, pages 6–28, New Orleans, Louisiana. https://eric.ed.gov/ ?id=ED376173. – Cité pages 35 et 38.

[Claveau et al., 2014] Claveau, V., Kijak, E., and Ferret, O. (2014). Explorer le graphe de voisinage pour améliorer les thésaurus distributionnels. In *Actes 21ème conférence sur le Traitement Automatique des Langues Naturelles*, Marseille, France. hal-01027787. – Cité page 59.

[Cohen et al., 2017] Cohen, K. B., Goss, F., Zweigenbaum, P., and Hunter, L. E. (2017). Translational morphosyntax: distribution of negation in clinical records and biomedical journal articles. *Studies in Health Technology and Informatics*, 245:346–350. doi:10.3233/978-1-61499-830-3-346. – Cité page 17.

[Collobert and Watson, 2008] Collobert, R. and Watson, J. (2008). A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings*

of the 25th international conference on Machine learning, pages 160–167, Helsinki, Finland. doi:10.1145/1390156.1390177. – Cité page 9.

[Cori, 2008] Cori, M. (2008). Des méthodes de traitement automatique aux linguistiques fondées sur les corpus. *Langages*, 3(171). doi:10.3917/lang.171.0095. – Cité page 7.

[Daud and Ali, 2018] Daud, N. and Ali, A. Z. M. (2018). The potentials of emoji in visual communication. *Idealogy*, 3(3):217–225. https://core.ac.uk/download/pdf/322854357.pdf. – Cité page 24.

[Deledalle, 1994] Deledalle, G. (1994). Charles S. Peirce. Les ruptures épistémologiques et les nouveaux paradigmes. *Travaux du Centre de Recherches Sémiologiques*, 62:51–66. http://-doc.rero.ch/record/255725. – Cité page 35.

[Deléger et al., 2015] Deléger, L., Grouin, C., and Bossy, R. (2015). Hybrid approaches for the DNER task at BioCreative V. In *Proceedings of the fifth BioCreative challenge evaluation workshop*, Sevilla, Spain. https://biocreative.bioinformatics.udel.edu/media/store/files/2015/BCV2015_paper_22.pdf. – Cité page 49.

[Deléger et al., 2010] Deléger, L., Grouin, C., and Zweigenbaum, P. (2010). Extracting medical information from narrative patient records: the case of medication-related information. *Journal of the American Medical Informatics Association*, 17(5):555–558. doi:10.1136/jamia.2010.003962. – Cité page 49.

[Delestre and Amar, 2022] Delestre, C. and Amar, A. (2022). DistilCamemBERT : une distillation du modèle français CamemBERT. In *Actes de CAp (Conférence sur l'Apprentissage automatique)*, Vannes, France. hal-03674695. – Cité page 67.

[Dernoncourt et al., 2017] Dernoncourt, F., Lee, J. Y., and Szolovits, P. (2017). NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Copenhagen, Denmark. doi:10.18653/v1/D17-2017. – Cité page 56.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. doi:10.18653/v1/N19-1423. – Cité page 66.

[Dib et al., 2022] Dib, F., Mayaud, P., Chauvin, P., and Launay, O. (2022). Online mis/disinformation and vaccine hesitancy in the era of COVID-19: why we need an ehealth literacy revolution. *Human vaccines & immunotherapeutics*, 18(1):1–3. doi:10.1080/21645515.2021.1874218. – Cité page 15.

[do Amaral et al., 2000] do Amaral, M. B., Roberts, A., and Rector, A. L. (2000). NLP techniques associated with the OpenGALEN ontology for semi-automatic textual extraction of medical knowledge: abstracting and mapping equivalent linguistic and logical constructs. In *Proceedings of the AMIA Symposium*, pages 76–80. PMID:11079848. – Cité page 63.

[Doucy and Massoussi, 2012] Doucy, G. and Massoussi, T. (2012). Sémantique inférentielle et compréhension des verbatim clients. In *Congrès Mondial de Linguistique Française*, pages 859–869. SHS Web of Conferences. doi:10.1051/shsconf/20120100230. – Cité page 36.

[Dumez, 2012] Dumez, H. (2012). Qu'est-ce que l'abduction, et en quoi peut-elle avoir un rapport avec la recherche qualitative ? *Le Libellio*, 8(3):3–9. http://lelibellio.com/wp-content/uploads/2013/01/Libellio27.pdf. – Cité pages 35 et 38.

[El Boukkouri, 2020] El Boukkouri, H. (2020). Ré-entraîner ou entraîner soi-même ? Stratégies de pré-entraînement de BERT en domaine médical. In *Actes des Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)*,

pages 29–42, Nancy, France. ATALA et AFCP. https://aclanthology.org/2020.jeptalnrecital-recital.3/. – Cité pages 66 et 72.

[El Boukkouri, 2021] El Boukkouri, H. (2021). *Domain Adaptation of Word Embeddings Through the Exploitation of In-domain Corpora and Knowledge Bases*. Thèse de doctorat, spécialité informatique, Université Paris-Saclay. tel-03560502v1. – Cité page 69.

[El Boukkouri et al., 2021] El Boukkouri, H., Ferret, O., Lavergne, T., Noji, H., Zweigenbaum, P., and Tsujii, J. (2021). CharacterBERT: reconciling ELMo and BERT for word-level open-vocabulary representations from characters. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6903–6915, Barcelona, Spain (Online). International Committee on Computational Linguistics. https://aclanthology.org/2020.coling-main.609/. – Cité page 69.

[El Boukkouri et al., 2019] El Boukkouri, H., Ferret, O., Lavergne, T., and Zweigenbaum, P. (2019). Embedding strategies for specialized domains: application to clinical entity recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics : Student Research Workshop*, pages 295–301, Florence, Italy. Association for Computational Linguistics. doi:10.18653/v1/P19-2041. – Cité page 67.

[Feng et al., 2020] Feng, Z., Guo, D., Tang, D., Duan, N., Feng, X., Gong, M., Shou, L., Qin, B., Liu, T., Jiang, D., and Zhou, M. (2020). CodeBERT: a pre-trained model for programming and natural languages. In *Findings of the Association for Computational Linguistics : EMNLP 2020*, pages 1536–1547, Online. Association for Computational Linguistics. doi:10.18653/v1/2020.findings-emnlp.139. – Cité page 66.

[Ferández-Gavilanes et al., 2018] Ferández-Gavilanes, M., Juncal-Martínez, J., García-Méndez, S., Costa-Montenegro, E., and no, F. J. G.-C. (2018). Creating emoji lexica from unsupervised sentiment analysis of their descriptions. *Expert Systems with Applications*, 103(1):74–91. doi:10.1016/j.eswa.2018.02.043. – Cité page 25.

[Ferret, 2015] Ferret, O. (2015). Déclasser les voisins non sémantiques pour améliorer les thésaurus distributionnels. In *Actes TALN*, Caen, France. cea-01858465. – Cité page 59.

[Friedman et al., 2006] Friedman, C., Borlawsky, T., Shagina, L., Xing, H. R., and Lussier, Y. A. (2006). Bio-ontology and text : bridging the modeling gap. *Bioinformatics*, 22(19). doi:10.1093/bioinformatics/btl405. – Cité page 63.

[Friedman et al., 2002] Friedman, C., Kra, P., and Rzhetsky, A. (2002). Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, 35(4):222–235. doi:10.1016/s1532-0464(03)00012-1. – Cité page 17.

[Gage, 1994] Gage, P. (1994). A new algorithm for data compression. *The C Users Journal*, 12(2):23–38. acm:10.5555/177910.177914. – Cité page 68.

[Gallant, 1991] Gallant, S. I. (1991). A practical approach for representing context and for performing word sense disambiguation using neural networks. *Neural Computation*, 3(3):293–309. doi:10.1162/neco.1991.3.3.293. – Cité page 59.

[Ghinassi, 2021] Ghinassi, I. (2021). Unsupervised text segmentation via deep sentence encoders: a first step towards a common framework for text-based segmentation, summarization and indexing of media content. unsupervised text segmentation via deep sentence encoders. In *2nd International Workshop on Data-driven Personalisation of Television (DataTV-2021)*, New York, NY. doi:10.5281/zenodo.4744399. – Cité page 79.

[Goeuriot et al., 2015] Goeuriot, L., Kelly, L., Suominen, H., Hanlen, L., Névéol, A., Grouin, C., Palotti, J., and Zuccon, G. (2015). Overview of the CLEF eHealth evaluation lab 2015. In *CLEF 2015 - 6th Conference and Labs of the Evaluation Forum*, Toulouse, France. Lecture Notes in Computer Science (LNCS), Springer. doi:10.1007/978-3-319-24027-5_44. – Cité page 50.

[Grabar and Grouin, 2022] Grabar, N. and Grouin, C. (2022). Year 2021: COVID-19, information extraction and BERTization among the most hot topics in medical natural language processing. *Yearbook of Medical Informatics*, 31(01). à paraître. – Cité page 67.

[Grabar et al., 2019] Grabar, N., Grouin, C., Hamon, T., and Claveau, V. (2019). Recherche et extraction d'information dans des cas cliniques. présentation de la campagne d'évaluation DEFT 2019. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Défi Fouille de Textes (atelier TALN-RECITAL)*, pages 7–16, Toulouse, France. hal-02280852. – Cité page 50.

[Grèzes et al., 2021] Grèzes, F., Blanco-Cuaresma, S., Accomazzi, A., Kurtz, M. J., Shapurian, G., Henneken, E. A., Grant, C. S., Thompson, D. M., Chyla, R., McDonald, S., Hostetler, T. W., Templeton, M. R., Lockhart, K. E., Martinovic, N., Chen, S., Tanner, C., and Protopapas, P. (2021). Building astroBERT, a language model for astronomy & astrophysics. *CoRR*, abs/2112.00590. arxiv:2112.00590. – Cité pages 66 et 72.

[Gross, 1981] Gross, M. (1981). Les bases empiriques de la notion de prédicat sémantique. *Langages*, 63:7–52. https://www.persee.fr/doc/lgge_0458-726x_1981_num_15_63_1875. – Cité page 11.

[Grouin, 2013a] Grouin, C. (2013a). *Anonymisation de documents cliniques : performances et limites des méthodes symboliques et par apprentissage statistique*. Thèse de doctorat, spécialité informatique biomédicale, Université Pierre et Marie Curie, Paris, France. tel-00848672. – Cité page 49.

[Grouin, 2013b] Grouin, C. (2013b). Building a contrasting taxa extractor for relation identification from assertions: biological taxonomy & ontology phrase extraction system. In *BioNLP-ST Workshop Proceedings*, pages 144–152, Sofia, Bulgaria. Association for Computational Linguistics. https://aclanthology.org/W13-2022. – Cité page 49.

[Grouin, 2014a] Grouin, C. (2014a). Clinical records de-identification using CRF and rule-based approaches. In *i2b2/UTHealth Shared-Tasks Proc*, Washington, DC. – Cité page 49.

[Grouin, 2014b] Grouin, C. (2014b). Identification of medication side effects in clinical records: an experiment based on the 2014 i2b2/uthealth corpus. In *i2b2/UTHealth Shared-Tasks Proc*, Washington, DC. – Cité page 49.

[Grouin, 2016a] Grouin, C. (2016a). Identification of mentions and relations between bacteria and biotope from PubMed abstracts. In *Proceedings of the BioNLP-ST*, pages 64–72, Berlin, Germany. Association for Computational Linguistics. doi:10.18653/v1/W16-3008. – Cité page 49.

[Grouin, 2016b] Grouin, C. (2016b). LIMSI at CEGS N-GRID 2016 NLP Shared-Tasks: track 1.a de-identification of unseen clinical texts. In *Proceedings of the CEGS N-GRID Work*, Chicago, IL. – Cité page 49.

[Grouin, 2016c] Grouin, C. (2016c). LIMSI at CEGS N-GRID 2016 NLP Shared-Tasks: track 1.b de-identification of clinical texts at character and token levels. In *Proceedings of hte CEGS N-GRID Work*, Chicago, IL. – Cité page 49.

[Grouin, 2018] Grouin, C. (2018). Simplification de schémas d'annotation : un aller sans retour ? In *Actes de la Conférence TALN. Volume 1 — Articles longs, articles courts de TALN*, pages 481–488, Rennes, France. Association pour le Traitement Automatique des Langues. hal-01831221. – Cité pages 9 et 55.

[Grouin, 2022] Grouin, C. (2022). Impact du français inclusif sur les outils du TAL. In Estève, Y., Jiménez, T., Parcollet, T., and Zanon Boito, M., editors, *Traitement Automatique des Langues Naturelles*, pages 126–135, Avignon, France. ATALA. hal-03704005. – Cité pages 10 et 27.

[Grouin et al., 2011a] Grouin, C., Dinarelli, M., Rosset, S., Wisniewski, G., and Zweigenbaum, P. (2011a). Coreference resolution in clinical reports. the LIMSI participation in the i2b2/va 2011 challenge. In *i2b2/VA Workshop Proc*, Washington, DC. – Cité page 49.

[Grouin et al., 2011b] Grouin, C., Forest, D., Paroubek, P., and Zweigenbaum, P. (2011b). Présentation et résultats du défi fouille de texte DEFT2011. quand un article de presse a-t-il été écrit ? à quel article scientifique correspond ce résumé ? In *Actes de DEFT*, Montpellier, France. TALN. – Cité page 17.

[Grouin et al., 2013a] Grouin, C., Grabar, N., Hamon, T., Rosset, S., Tannier, X., and Zweigenbaum, P. (2013a). Eventual situations for timeline extraction from clinical reports. *Journal of the American Medical Informatics Association*, 20(5):820–827. doi:10.1136/amiajnl-2013-001627. – Cité page 49.

[Grouin et al., 2021] Grouin, C., Grabar, N., and Illouz, G. (2021). Classification de cas cliniques et évaluation automatique de réponses d'étudiants : présentation de la campagne DEFT 2021. In *Actes de l'atelier Défi Fouille de Textes@TALN 2021 Classification de cas cliniques et correction automatique de copies d'étudiants. Atelier DÉfi Fouille de Textes*, pages 1–13, Lille, France. Association pour le Traitement Automatique des Langues. hal-03265926. – Cité page 50.

[Grouin et al., 2014a] Grouin, C., Megahed, D., and Zweigenbaum, P. (2014a). Medication side effects identification from clinical records and health social media. In *Forum STIC Paris-Saclay*, Palaiseau, France. https://perso.limsi.fr/grouin/supports/2014_stic_grouin.pdf. – Cité page 52.

[Grouin and Moriceau, 2016] Grouin, C. and Moriceau, V. (2016). LIMSI at SemEval-2016 task 12: machine-learning and temporal information to identify clinical events and time expressions. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, CA. Association for Computational Linguistics. doi:10.18653/v1/S16-1190. – Cité pages 23 et 50.

[Grouin et al., 2014b] Grouin, C., Moriceau, V., Rosset, S., and Zweigenbaum, P. (2014b). Risk factor identification from clinical records for diabetic patients. In *i2b2/UTHealth Shared-Tasks Proc*, Washington, DC. – Cité page 49.

[Grouin et al., 2015] Grouin, C., Moriceau, V., and Zweigenbaum, P. (2015). Combining glass box and black box evaluations in the identification of heart disease risk factors and their temporal relations from clinical records. *J Biomed Inform*, 58:S133–S142. doi:10.1016/j.jbi.2015.06.014. – Cité page 49.

[Grouin et al., 2013b] Grouin, C., Paroubek, P., and Zweigenbaum, P. (2013b). DEFT2013 se met à table présentation du défi et résultats. In *Actes de DEFT*, Les Sables-d'Olonnes, France. TALN. – Cité page 17.

[Gu et al., 2022] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2022). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23. doi:10.1145/3458754. – Cité page 66.

[Guespin, 1971] Guespin, L. (1971). Problématique des travaux sur le discours politique. *Langages*, 23:3–24. doi:10.3406/lgge.1971.2048. – Cité page 33.

[Guibon et al., 2016] Guibon, G., Ochs, M., and Bellot, P. (2016). From emojis to sentiment analysis. In *Proceedings of the Workshop Affect Compagnon Artificiel Interaction*, Brest, France. hal-01529708. – Cité page 25.

[Hamon et al., 2015] Hamon, T., Fraisse, A., Paroubek, P., Zweigenbaum, P., and Grouin, C. (2015). Analyse des émotions, sentiments et opinions exprimés dans les tweets : présentation et résultats de l'édition 2015 du défi fouille de texte (DEFT). In *Actes de 11e Défi Fouille de Texte*

(DEFT'2015), Caen, France. TALN, Association pour le Traitement Automatique des Langues. http://talnarchives.atala.org/ateliers/2015/DEFT/deft-2015-long-001.pdf. – Cité page 16.

[Hamon et al., 2014a] Hamon, T., Grouin, C., and Zweigenbaum, P. (2014a). Disease and disorder template filling using rule-based and statistical approaches. In *Proc of ShARe/CLEF eHealth Evaluation Lab*, Sheffield, UK. hal-01831232. – Cité page 49.

[Hamon et al., 2014b] Hamon, T., Pleplé, Q., Paroubek, P., Zweigenbaum, P., and Grouin, C. (2014b). Analyse automatique de textes littéraires et scientifiques : présentation et résultats du défi fouille de texte DEFT2014. In *Actes de Atelier DÉfi Fouille de Textes à TALN 2014 (DEFT'2014)*, Marseille, France. TALN, Association pour le Traitement Automatique des Langues. http://talnarchives.atala.org/ateliers/2014/DEFT/1.pdf. – Cité page 17.

[Harris, 1952] Harris, Z. S. (1952). Discourse Analysis. *Language*, 28(1):1–30. doi:10.2307/409987. – Cité page 37.

[Harris, 1954] Harris, Z. S. (1954). Distributional Structure. *Word*, 10(23):146–162. doi:10.1080/00437956.1954.11659520. – Cité page 8.

[Hassan et al., 2018] Hassan, N. Y., Gomaa, W. H., Khoriba, G. A., and Haggag, M. H. (2018). Supervised learning approach for Twitter credibility detection. In *Proceedings of the 13th International Conference on Computer Engineering and Systems (ICCES)*, pages 196–201. doi:10.1109/icces.2018.8639315. – Cité page 15.

[HCE, 2015] HCE (2015). *Pour une communication publique sans stéréotype de sexe*. Haut Conseil à l'Égalité entre les femmes et les hommes. http://bit.ly/2fejwZ7. – Cité page 27.

[Hearst, 1997] Hearst, M. A. (1997). Text Tiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64. https://aclanthology.org/J97-1003. – Cité page 79.

[Heiden, 2010] Heiden, S. (2010). The TXM Platform: building open-source textual analysis software compatible with the TEI encoding scheme. In *Proc of Pacific Asia Conference on Language, Information and Computation*, Sendai, Japon. https://halshs.archives-ouvertes.fr/-docs/00/54/97/64/PDF/paclic24_sheiden.pdf. – Cité page 39.

[Heiden et al., 2010] Heiden, S., Magué, J.-P., and Pincemin, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie — conception et développement. In *Proc of 10th International Conference on the Statistical Analysis of Textual Data (JADT)*, Rome, Italie. https://halshs.archives-ouvertes.fr/docs/00/54/97/79/PDF/Heiden_al_jadt2010.pdf. – Cité page 39.

[Henry et al., 2018] Henry, D., Stattner, E., and Collard, M. (2018). Filter hashtag context through an original data cleaning method. *Procedia Computer Science*, 130:464–471. doi:10.1016/j.procs.2018.04.050. – Cité page 26.

[Hinton et al., 2014] Hinton, G., Vinyals, O., and Dean, J. (2014). Distilling the knowledge in a neural network. In *Proceedings of the NIPS 2014 Deep Learning Workshop*, Montréal, Canada. doi:10.48550/arXiv.1503.02531. – Cité page 67.

[Hromadova, 2019] Hromadova, C. (2019). Des bleus à l'âme ou l'accident littéraire chez Françoise Sagan. *Crossways Journal*, 2(3). https://crossways.lib.uoguelph.ca/index.php/crossways/article/download/5395/5154. – Cité page 20.

[Jakobson, 1963] Jakobson, R. (1963). Les embrayeurs, les catégories verbales et le verbe russe. In *Les Fondations du langage. Essais de linguistique générale I*, chapter 8. Éditions de Minuit, Paris. – Cité page 33.

[Ji et al., 2021] Ji, Y., Zhou, Z., Liu, H., and Davuluri, R. V. (2021). DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120. doi:10.1093/bioinformatics/btab083. – Cité page 66.

[Jiang et al., 2007] Jiang, L., Zhou, M., Chien, L.-F., and Niu, C. (2007). Named entity translation with web mining and transliteration. In *Proceedings of the 20th international joint conference on Artifical intelligence (IJCAI)*, pages 1629–1634, Hyderabad, India. https://www.ijcai.org/-Proceedings/07/Papers/263.pdf. – Cité page 19.

[Johnson et al., 2016] Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3. doi:10.1038/sdata.2016.35. – Cité page 69.

[Kaewphan et al., 2014] Kaewphan, S., Hakaka, K., and Ginter, F. (2014). UTU: disease mention recognition and normalization with CRFs and vector space representations. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 807–811, Dublin, Ireland. doi:10.3115/v1/S14-2143. – Cité page 60.

[Kamal Eddine et al., 2021] Kamal Eddine, M., Tixier, A., and Vazirgiannis, M. (2021). BARThez: a skilled pretrained French sequence-to-sequence model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9369–9390, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. doi:10.18653/v1/2021.emnlp-main.740. – Cité page 67.

[Kant, 1781] Kant, E. (1781). *Critique de la raison pure*. Kritik der reinen Vernunft. – Cité page 35.

[Karapetiantz et al., 2018] Karapetiantz, P., Bellet, F., Audeh, B., Lardon, J., Leprovost, D., Aboukhamis, R., Morlane-Hondère, F., Grouin, C., Burgun, A., Katsahian, S., Jaulent, M.-C., Beyens, M.-N., Lillo-Le Louët, A., and Bousquet, C. (2018). Descriptions of adverse drug reactions are less informative in forums than in the French pharmacovigilance database but provide more unexpected reactions. *Frontiers in Pharmacology*, 9. doi:10.3389/fphar.2018.00439. – Cité page 51.

[Karoui et al., 2017] Karoui, J., Benamara, F., Moriceau, V., Patti, V., Bosco, C., and Aussenac-Gilles, N. (2017). Exploring the impact of pragmatic phenomena on irony detection in Tweets: a multilingual corpus study. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 262–272, Valencia, Spain. Association for Computational Linguistics. https://aclanthology.org/E17-1025. – Cité page 25.

[Kudo and Richardson, 2018] Kudo, T. and Richardson, J. (2018). SentencePiece: a simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics. doi:10.18653/v1/D18-2012. – Cité page 68.

[Kupść, 2008] Kupść, A. (2008). Adjectives in treelex. In *16th International Conference Intelligent Information Systems*, pages 287–296, Zakopane, Poland. Academic Publishing House. http://redac.univ-tlse2.fr/lexiques/treelex/adj_iis_final_all.pdf. – Cité page 80.

[Kupść and Abeillé, 2008] Kupść, A. and Abeillé, A. (2008). Growing TreeLex. In *9th International Conference, CICLing*, Haïfa, Israël. hal:inria-00338103. – Cité page 80.

[Labov, 1972] Labov, W. (1972). *Sociolinguistic patterns*. University of Pennsylvania Press, Philadelphia. – Cité page 8.

[Labutov and Lipson, 2013] Labutov, I. and Lipson, H. (2013). Re-embedding words. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 489–493, Sofia, Bulgaria. https://aclanthology.org/P13-2087. – Cité page 67.

[Lan et al., 2019] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). ALBERT : A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942. https://openreview.net/pdf?id=H1eA7AEtvS. – Cité page 67.

[Lauscher et al., 2020] Lauscher, A., Ravishankar, V., Vulić, I., and Glavaš, G. (2020). From zero to Hero: on the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics. doi:10.18653/v1/2020.emnlp-main.363. – Cité page 85.

[Lavergne et al., 2010] Lavergne, T., Cappé, O., and Yvon, F. (2010). Practical very large scale CRFs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 504–513, Uppsala, Sweden. https://aclanthology.org/P10-1052/. – Cité pages 53 et 58.

[Le et al., 2020a] Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2020a). FlauBERT : des modèles de langue contextualisés pré-entraînés pour le français. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles. Volume 2 : Traitement Automatique des Langues Naturelles (Articles courts)*, pages 268–278, Nancy, France. Association pour le Traitement Automatique des Langues. hal-02784776v3. – Cité page 66.

[Le et al., 2020b] Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2020b). FlauBERT: unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association. hal-02890258. – Cité page 66.

[Lee et al., 2019] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240. doi:10.1093/bioinformatics/btz682. – Cité pages 66, 67 et 72.

[Lewis et al., 2020] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.703. – Cité page 67.

[Li et al., 2021] Li, D., Xiong, Y., Hu, B., Tang, B., Peng, W., and Chen, Q. (2021). Drug knowledge discovery via multi-task learning and pre-trained models. *BMC Medical Informatics in Decision Making*, 21(9):251. doi:10.1186/s12911-021-01614-7. – Cité page 63.

[Lindberg et al., 1993] Lindberg, D. A., Humphreys, B. L., and McRay, A. T. (1993). The Unified Medical Language System. *Methods of Information in Medicine*, 32(4):281–291. doi:10.1055/s-0038-1634945. – Cité page 60.

[Liu et al., 2021] Liu, C., Fang, F., Lin, X., Cai, T., Tan, X., Liu, J., and Lu, X. (2021). Improving sentiment analysis accuracy with emoji embedding. *Journal of Safety Science and Resilience*, 2(4):246–252. doi:10.1016/j.jnlssr.2021.10.003. – Cité page 25.

[Liu et al., 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: a robustly optimized BERT pretraining approach. doi:10.48550/arxiv.1907.11692. – Cité page 66.

[Lorig et al., 2001] Lorig, K. R., Sobel, D. S., Ritter, P. L., Laurent, D., and Hobbs, M. (2001). Effect of a self-management program on patients with chronic disease. *Effective Clinical Practice*, 4(6):256–262. pmid:11769298. – Cité page 15.

[Mahajan et al., 2016] Mahajan, D., Kolathur, V., Bansal, C., Parthasarathy, S., Sellamanickam, S., Keerthi, S., and Gehrke, J. (2016). Hashtag recommendation for enterprise applications. In for Computing Machinery, A., editor, *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 893–902, Indianapolis, USA. doi:10.1145/2983323.2983365. – Cité page 26.

[Maingueneau, 1979] Maingueneau, D. (1979). L'analyse du discours. *Repères*, 51:3–27. doi:10.3406/reper.1979.1614. – Cité page 33.

[Marcandier, 2011] Marcandier, C. (2011). *Le théâtre*, chapter 17. Nathan. hal-01722125. – Cité page 73.

[Marchand, 2015] Marchand, M. (2015). *Domaines et fouille d'opinion : une étude des marqueurs multi-polaires au niveau du texte*. Thèse de doctorat, spécialité informatique, Université Paris-Sud. tel-01157951. – Cité page 72.

[Marchello-Nizia, 1989] Marchello-Nizia, C. (1989). Le neutre et l'impersonnel. *Linx*, 1(21):173–179. Genre et langage. Actes du colloque tenu à Paris X-Nanterre les 14-15-16 décembre 1988. doi:10.3406/linx.1989.1139. – Cité page 30.

[Martin et al., 2020] Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., and Sagot, B. (2020). CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.645. – Cité page 66.

[Megahed, 2014] Megahed, D. (2014). Etude des forums de santé pour la détection d'événements secondaires. Mémoire de master ingénierie linguistique, Institut National des Langues et Civilisations Orientales, Paris. – Cité page 52.

[Mikolov et al., 2013] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances on Neural Information Processing Systems*, NIPS'13, pages 3111— 3119, Red Hook, NY. Curran Associates Inc. acm:10.5555/2999792.2999959. – Cité pages 59 et 65.

[Minard et al., 2011] Minard, A.-L., Ligozat, A.-L., Ben Abacha, A., Bernhard, D., Cartoni, B., Deléger, L., Grau, B., Rosset, S., Zweigenbaum, P., and Grouin, C. (2011). Hybrid methods for improving information access in clinical documents : concept, assertion, and relation identification. *J Am Med Inform Assoc*, 18:588–93. doi:10.1136/amiajnl-2011-000154. – Cité pages 23 et 49.

[Mintz et al., 2009] Mintz, M., Bills, S., Snow, R., and 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*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics. https://aclanthology.org/P09-1113. – Cité page 61.

[Moeschler, 2017] Moeschler, J. (2017). Pragmatique du discours. In Pavelin Lesic, B., editor, *Francontraste*, pages 217–230, Mons. http://archive-ouverte.unige.ch/unige:110179. – Cité page 33.

[Moreau, 2019] Moreau, M.-L. (2019). L'accord de proximité dans l'écriture inclusive. peut-on utiliser n'importe quel argument ? In Dister, A. and Piron, S., editors, *Les discours de référence sur la langue française*, pages 351–378. Presses de l'Université Saint-Louis, Bruxelles, Belgique. doi:10.4000/books.pusl.26517. – Cité page 27.

[Morlane-Hondère and Grouin, 2016] Morlane-Hondère, F. and Grouin, C. (2016). Normalisation de concepts cliniques par des vecteurs de mots. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2016. Volume 7 : COLTAL*, pages 27–35, Paris, France. Association pour le Traitement Automatique des Langues. hal-01831234. – Cité pages 10 et 60.

[Morlane-Hondère et al., 2015a] Morlane-Hondère, F., Grouin, C., Moriceau, V., and Zweigenbaum, P. (2015a). Médicaments qui soignent, médicaments qui rendent malade : étude des relations causales pour identifier les effets secondaires. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles*, Caen, France. Association pour le Traitement Automatique des Langues. hal-02950996. – Cité pages 10 et 58.

[Morlane-Hondère et al., 2015b] Morlane-Hondère, F., Grouin, C., and Zweigenbaum, P. (2015b). étude des verbes introducteurs de noms de médicaments dans les forums de santé. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 21–27, Caen, France. ATALA. https://aclanthology.org/2015.jeptalnrecital-court.4/. – Cité pages 10 et 59.

[Morlane-Hondère et al., 2016a] Morlane-Hondère, F., Grouin, C., and Zweigenbaum, P. (2016a). Identification of drug-related medical conditions in social media. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia. European Language Resources Association (ELRA). https://aclanthology.org/L16-1320. – Cité pages 10, 54, 57 et 59.

[Morlane-Hondère et al., 2016b] Morlane-Hondère, F., Grouin, C., and Zweigenbaum, P. (2016b). Représentation des informations textuelles pour la détection d'états pathologiques par apprentissage statistique. In Lovis, C., Séroussi, B., Ugon, A., and Randriambelonoro, M. M., editors, *Journées Francophones d'informatique Médicale (JFIM)*, Genève. hal-01831156. – Cité pages 9 et 54.

[Mounin, 2004] Mounin, G. (2004). *Dictionnaire de la linguistique*. Presses Universitaires de France. – Cité page 30.

[Nangia et al., 2020] Nangia, N., Vania, C., Bhalerao, R., and Bowman, S. R. (2020). CrowS-Pairs: a challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online symposium. Association for Computational Linguistics. doi:10.18653/v1/2020.emnlp-main.154. – Cité page 72.

[Navigli, 2009] Navigli, R. (2009). Word sense disambiguation: a survey. *ACM Computing Surveys*, 41(2). doi:10.1145/1459352.1459355. – Cité page 59.

[Nayak et al., 2020] Nayak, A., Timmapathini, H., Ponnalagu, K., and Gopalan Venkoparao, V. (2020). Domain adaptation challenges of BERT in tokenization and sub-word representations of out-of-vocabulary words. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 1–5, Online symposium. Association for Computational Linguistics. doi:10.18653/v1/2020.insights-1.1. – Cité page 71.

[Neff and Nagy, 2016] Neff, G. and Nagy, P. (2016). Talking to bots: symbiotic agency and the case of Tay. *International Journal of Communication*, 10:4915–4931. https://ijoc.org/index.php/ijoc/article/view/6277. – Cité page 73.

[Névéol et al., 2017a] Névéol, A., Anderson, R. N., Cohen, K. B., Grouin, C., Lavergne, T., Rey, G., Robert, A., Rondet, C., and Zweigenbaum, P. (2017a). ICD-10 coding of death certificates in multiple languages: the CLEF eHealth 2016 and 2017 shared tasks. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum*, Dublin, Ireland. http://ceur-ws.org/Vol-1866/invited_paper_6.pdf. – Cité page 50.

[Névéol et al., 2017b] Névéol, A., Anderson, R. N., Cohen, K. B., Grouin, C., Lavergne, T., Rey, G., Robert, A., and Zweigenbaum, P. (2017b). CLEF eHealth 2017 multilingual information extraction task overview: ICD10 coding of death certificates in English and French. In *Proceedings of the CLEF eHealth Evaluation lab*, Dublin, Ireland. hal-01665374. – Cité page 50.

[Névéol et al., 2022] Névéol, A., Dupont, Y., Bezançon, J., and Fort, K. (2022). French CrowS-Pairs: extending a challenge dataset for measuring social bias in masked language models to a language other than English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531, Dublin, Ireland. Association for Computational Linguistics. doi:10.18653/v1/2022.acl-long.583. – Cité page 72.

[Névéol et al., 2016] Névéol, A., Grouin, C., Cohen, K. B., Hamon, T., Lavergne, T., Kelly, L., Goeuriot, L., Rey, G., Robert, A., Tannier, X., and Zweigenbaum, P. (2016). Clinical information extraction at the CLEF eHealth evaluation lab 2016. In *Proceedings of the CLEF eHealth Evaluation lab*, Evora, Portugal. hal-01922402. – Cité page 50.

[Névéol et al., 2015] Névéol, A., Grouin, C., Tannier, X., Hamon, T., Kelly, L., Goeuriot, L., and Zweigenbaum, P. (2015). CLEF eHealth evaluation lab 2015 task 1b: clinical named entity recognition. In *Proceedings of the ShARe/CLEF Evaluation Lab*, Toulouse, France. hal-01922444. – Cité page 50.

[Nguyen et al., 2020] Nguyen, D. Q., Vu, T., and Tuan Nguyen, A. (2020). BERTweet: a pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics. doi:10.18653/v1/2020.emnlp-demos.2. – Cité page 66.

[Oprean et al., 2014] Oprean, C., Mokbel, C., Likforman-Sulem, L., and Popescu, A. (2014). Reconnaissance de mots manuscrits hors-vocabulaire en utilisant des ressources web. *Document numérique*, 17(3):77–96. https://www.cairn.info/revue-document-numerique-2014-3-page-77.htm. – Cité page 68.

[O'Reilly, 2005] O'Reilly, T. (2005). What is web 2.0. https://www.oreilly.com/pub/a//-web2/archive/what-is-web-20.html. – Cité page 15.

[Park et al., 2018] Park, J. H., Shin, J., and Fung, P. (2018). Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics. doi:10.18653/v1/D18-1302. – Cité page 72.

[Paroubek, 2013] Paroubek, P. (2013). *De l'évaluation en Traitement Automatique des Langues*. Habilitation à diriger des recherches, Université Paris-Sud. https://perso.limsi.fr/pap/hdr_memoir_pap.pdf. – Cité page 7.

[Paroubek et al., 2018] Paroubek, P., Grouin, C., Bellot, P., Claveau, V., Eshkol-Taravella, I., Fraisse, A., Jackiewicz, A., Karoui, J., Monceaux, L., and Torres-Moreno, J.-M. (2018). Deft2018 : recherche d'information et analyse de sentiments dans des tweets concernant les transports en île de france. In *Actes de la 25ème Conférence sur le Traitement Automatique des Langues Naturelles (DEFT'2018)*, Rennes, France. Association pour le Traitement Automatique des Langues. http://talnarchives.atala.org/ateliers/2018/DEFT/1.pdf. – Cité page 16.

[Paroubek et al., 2012] Paroubek, P., Zweigenbaum, P., Forest, D., and Grouin, C. (2012). Indexation libre et contrôlée d'articles scientifiques. présentation et résultats du défi fouille de textes DEFT2012. In *Actes de DEFT*, Grenoble, France. TALN. – Cité page 17.

[Peirce, 1878] Peirce, C. S. (1878). Deduction, induction, and hypothesis. *Popular Science Monthly*, 13:470–482. Illustrations of the Logic of Science VI. – Cité page 34.

[Peng et al., 2019] Peng, Y., Yan, S., and Lu, Z. (2019). Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*. doi:10.48550/arXiv.1906.05474. – Cité page 66.

[Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics. doi:10.3115/v1/D14-1162. – Cité page 65.

[Périnet and Hamon, 2014] Périnet, A. and Hamon, T. (2014). Réduction de la dispersion des données par généralisation des contextes distributionnels : application aux textes de spécialité. In *Actes Conférence sur le Traitement Automatique des Langues Naturelles*, Marseille, France. hal-01972768. – Cité page 59.

[Peters et al., 2018] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and 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)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics. doi:10.18653/v1/N18-1202. – Cité page 65.

[Piolat and Bannour, 2009] Piolat, A. and Bannour, R. (2009). Emotaix : un scénario de Tropes pour l'identification automatisée du lexique émotionnel et affectif. *L'Année psychologique*, 109(4):655–698. http://centrepsycle-amu.fr/wp-content/uploads/2014/01/Piolat_Bannour_2009_EMOTAIX.pdf. – Cité page 20.

[Podder et al., 2021] Podder, V., Lew, V., and Ghassemzadeh, S. (2021). SOAP notes. *StatPearls*. PMID:29489268. – Cité page 49.

[Poerner et al., 2020] Poerner, N., Waltinger, U., and Schütze, H. (2020). Inexpensive domain adaptation of pretrained language Models: case studies on biomedical NER and Covid-19 QA. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1482–1490, Online. Association for Computational Linguistics. doi:10.18653/v1/2020.findings-emnlp.134. – Cité pages 67 et 85.

[Polignano et al., 2019] Polignano, M., Basile, P., de Gemmis, M., Semeraro, G., and Basile, V. (2019). AlBERTo: Italian BERT language understanding model for NLP challenging tasks based on tweets. *Italian Journal of Computational Linguistics*, 5(2):11–31. doi:10.4000/ijcol.472. – Cité page 67.

[Poon et al., 2014] Poon, H., Toutanova, K., and Quirk, C. (2014). Distant supervision for cancer pathway extraction from text. In *Pacific Symposium on Biocomputing 2015*, pages 120–131, Kohala Coast, Hawaii. doi:10.1142/9789814644730_0013. – Cité page 61.

[Raithel et al., 2022] Raithel, L., Mutinda, F. W., Nishiyama, T., Laï-King, M., Yada, S., Roller, R., Grouin, C., Savary, A., Névéol, A., Lavergne, T., Aramaki, E., Möller, S., Matsumoto, Y., and Zweigenbaum, P. (2022). KEEPHA at n2c2 2022: track 1. In *Proceedings of the n2c2 NLP Challenge*, Washington, DC. – Cité page 49.

[Rastier, 1996] Rastier, F. (1996). La sémantique des textes : concepts et applications. *Hermes*, 16:15–37. http://www.revue-texto.net/Inedits/Rastier/Rastier_Concepts.html. – Cité pages 9 et 34.

[Rastier, 2011] Rastier, F. (2011). *La mesure et le grain. Sémantique de corpus*. Honoré Champion, Paris. Collection "Lettres numériques". – Cité page 8.

[Raymondie and Steiner, 2020] Raymondie, R. A. and Steiner, D. D. (2020). Stéréotypes de genre concernant l'expression des émotions : pensez subordonné–pensez femme ? In Lagabrielle, C., Steiner, D., and Battistelli, A., editors, *Carrières, leadership et conflits*, pages 203–217. Editions L'Harmattan. hal-02877275. – Cité page 74.

[Reichherzer et al., 2021] Reichherzer, P., Schüssler, F., Lefranc, V., Yusafzai, A., Alkan, A. K., Ashkar, H., and Becker Tjus, J. (2021). Astro-COLIBRI–the COincidence LIBrary for Real-time Inquiry for multimessenger astrophysics. *The Astrophysical Journal Supplement Series*, 256(1). doi:10.3847/1538-4365/ac1517. – Cité page 16.

[Reimers and Gurevych, 2019] Reimers, N. and Gurevych, I. (2019). Sentence-BERT: sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics. doi:10.18653/v1/D19-1410. – Cité page 66.

[Ren and Guo, 2021] Ren, W. and Guo, Y. (2021). What is "Versailles Literature"?: humblebrags on Chinese social networking sites. *Journal of Pragmatics*, 184:185–195. doi:10.1016/j.pragma.2021.08.002. – Cité page 20.

[Riban and Gerin, 2017] Riban, C. and Gerin, M. (2017). Les garçons et les filles sont belles. hal-01696511, vidéo sur https://www.youtube.com/watch ?v=8X45yYIF1Gw. – Cité page 27.

[Riedl and Biemann, 2012] Riedl, M. and Biemann, C. (2012). TopicTiling : A text segmentation algorithm based on LDA. In *Proceedings of ACL 2012 Student Research Workshop*, pages 37–42, Jeju Island, Korea. Association for Computational Linguistics. https://aclanthology.org/W12-3307. – Cité page 79.

[Roudaut, 2017] Roudaut, F. (2017). Comment on invente les hypothèses : Peirce et la théorie de l'abduction. *Cahiers phylosophiques*, 3(150):45–65. https://www.cairn.info/revue-cahiers-philosophiques-2017-3-page-45.htm. – Cité page 35.

[Roze, 2013] Roze, C. (2013). *Vers une algèbre des relations de discours*. Thèse de doctorat, spécialité sciences de l'homme et société, Université Paris-Diderot – Paris VII, Paris, France. tel-00881243. – Cité page 58.

[Roze et al., 2012] Roze, C., Danlos, L., and Muller, P. (2012). LEXCONN: a French lexicon of discourse connectives. *Discours*, 10. doi:10.4000/discours.8645. – Cité page 58.

[Rutledge, 1998] Rutledge, R. E. (1998). The Astronomer's Telegram: a web-based short-notice publication system for the professional astronomical community. *Publications of the Astronomical Society of the Pacific*, 110(748). doi:10.1086/316184. – Cité page 16.

[Sahlgren, 2008] Sahlgren, M. (2008). The distributional hypothesis. *Rivista di Linguistica*, 20(1):33–53. https://www.diva-portal.org/smash/get/diva2:1041938/FULLTEXT01.pdf. – Cité pages 8 et 72.

[Sanh et al., 2019] Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *Proceedings of the NeurIPS Workshop*. doi:10.48550/arxiv.1910.01108. – Cité page 67.

[Sarker et al., 2015] Sarker, A., Rachel, R., Nikfarjam, A., O'Connor, K., Smith, K., Jayaraman, S., Upadhaya, T., and Gonzales, G. (2015). Utilizing social media data for pharmacovigilance: a review. *Journal of Biomedical Informatics*, 54:202–212. doi:10.1016/j.jbi.2015.02.004 . – Cité page 54.

[Saussure, 1916] Saussure, F. (1916). *Cours de linguistique générale*. Payot, Lausanne, Paris. – Cité pages 8, 33 et 78.

[Schuster and Nakajima, 2012] Schuster, M. and Nakajima, K. (2012). Japanese and Korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152, Kyoto, Japan. doi:10.1109/icassp.2012.6289079. – Cité page 68.

[Sennrich et al., 2016] Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics. doi:10.18653/v1/P16-1162. – Cité page 68.

[Strubell et al., 2019] Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. doi:10.48550/arXiv.1906.02243. – Cité page 67.

[Stubbs et al., 2015] Stubbs, A., Kotfila, C., and Uzuner, O. (2015). Automated systems for the de-identification of longitudinal clinical narratives: overview of 2014 i2b2/UTHealth shared task track 1. *J Biomed Inform*, 58:S11–S19. doi:10.1016/j.jbi.2015.06.007. – Cité page 55.

[Sun et al., 2021] Sun, C., Yang, Z., Wang, L., Zhang, Y., Lin, H., and Wang, J. (2021). Biomedical named entity recognition using bert in the machine reading comprehension framework. *Journal of Biomedical Informatics*, 118:103799. doi:10.1016/j.jbi.2021.103799. – Cité page 66.

[Sun et al., 2019] Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., and Wang, W. Y. (2019). Mitigating gender bias in natural language processing: literature review. In *Proceedings of the 57th Annual Meeting of the Association for*

Computational Linguistics, pages 1630–1640, Florence, Italy. Association for Computational Linguistics. doi:10.18653/v1/P19-1159. – Cité page 72.

[Sundararaman et al., 2021] Sundararaman, D., Subramanian, V., Wang, G., Si, S., Shen, D., Wang, D., and Carin, L. (2021). Syntactic knowledge-infused transformer and BERT models. In *CEUR Wrokshop Proceedings*. ceur-ws.org/Vol-3052/short21.pdf. – Cité page 85.

[Tourrette, 2012] Tourrette, E. (2012). De l'égologie selon La Rochefoucauld. *Littérature*, 165(1):3–15. doi:10.3917/litt.165.0003. – Cité page 20.

[Uzuner et al., 2011] Uzuner, O., South, B. R., Shen, S., and DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc*, 18(5):552–556. doi:10.1136/amiajnl-2011-000203. – Cité page 22.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA. doi:10.48550/arXiv.1706.03762. – Cité page 65.

[Viennot, 2022] Viennot, E. (2022). De la parenthèse au point médian. des nouveaux habits de l'écriture inclusive et de la malhonnêteté de ses opposant·es. *Travail, genre et sociétés*, 1(47):165–168. doi:10.3917/tgs.047.0165. – Cité page 29.

[Vincze and Bestgen, 2011] Vincze, N. and Bestgen, Y. (2011). Identification de mots germes pour la construction d'un lexique de valence au moyen d'une procédure supervisée. In *Actes de la 18ème Conférence sur le Traitement Automatique des Langues Naturelles*, Montpellier, France. Association pour le Traitement Automatique des Langues. http://talnarchives.atala.org/TALN/TALN-2011/taln-2011-long-014.pdf. – Cité page 80.

[Wang et al., 2019] Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019). SuperGLUE: a stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, volume 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf. – Cité page 66.

[Wang et al., 2018] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). GLUE: a multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP : Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics. doi:10.18653/v1/W18-5446. – Cité page 66.

[Wang and Gan, 2017] Wang, L. and Gan, J. Q. (2017). Prediction of the 2017 French election based on Twitter data analysis. In *Proceedings of the 9th Computer Science and Electronic Engineering (CEEC)*, pages 89–93. doi:10.1109/ceec.2017.8101605. – Cité page 15.

[Wood, 2008] Wood, C. H. (2008). Time, cycles and tempos in social-ecological research and environmental policy. *Time & Society*, 17(2–3). doi:10.1177/0961463X08093425. – Cité page 83.

[Wu et al., 2022] Wu, L., Ali, S., Ali, H., Brock, T., Xu, J., and Tong, W. (2022). NeuroCORD: a language model to facilitate COVID-19-associated neurological disorder studies. *Int J Environ Res Public Health*, 19(16):9974. doi:10.3390/ijerph19169974. – Cité page 63.

[Wu et al., 2015] Wu, Y., Xu, J., Zhang, Y., and Xu, H. (2015). Clinical abbreviation disambiguation using neural word embeddings. In *Proceedings of BioNLP 15*, pages 171–176, Beijing, China. doi:10.18653/v1/W15-3822. – Cité page 59.

[Xu et al., 2008] Xu, L., Lin, H., Pan, Y., Ren, H., and Chen, J. (2008). Constructing the affective lexicon ontology. *Journal of the China Society for Scientific and Technical Information*, 27:180–185. – Cité page 42.

[Yan, 2018] Yan, L. (2018). Analyse des inférences pour la fouille d'opinion en chinois. In *Actes de la Conférence TALN. Volume 2 - Démonstrations, articles des Rencontres Jeunes Chercheurs, ateliers DeFT*, pages 17–26, Rennes, France. ATALA. hal-02507182. – Cité pages 9, 36 et 65.

[Yan, 2021] Yan, L. (2021). *Le rôle des inférences pour la fouille d'opinion : applications aux réseaux sociaux en langue chinoise.* Thèse de doctorat, spécialité science du langage, INaLCO. tel-03469568. – Cité pages 17 et 35.

[Yan et al., 2020] Yan, L., E, D., Gan, M., Grouin, C., and Valette, M. (2020). Inference annotation of a Chinese corpus for opinion mining. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4991–4999, Marseille, France. European Language Resources Association. hal-02507170. – Cité pages 9, 39 et 42.

[Ye et al., 2022] Ye, J., Wang, Z., and Hai, J. (2022). Social networking service, patient-generated health data, and population health informatics: national cross-sectional study of patterns and implications of leveraging digital technologies to support mental health and well-being. *Journal of Medical Internet Research*, 24(4):e30898. doi:10.2196/30898. – Cité page 15.

[Zhang et al., 2022] Zhang, K., Hu, B., Zhou, F., Song, Y., Zhao, X., and Huang, X. (2022). Graph-based structural knowledge-aware network for diagnosis assistant. *Math Biosci Eng*, 19(10):10533–10549. doi:10.3934/mbe.2022492. – Cité page 63.

[Zweigenbaum and Grabar, 2002] Zweigenbaum, P. and Grabar, N. (2002). Accentuation de mots inconnus : application au thesaurus biomédical MeSH. In *Actes de la 9ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, pages 53–62, Nancy, France. Association pour le Traitement Automatique des Langues. https://aclanthology.org/2002.jeptalnrecital-long.3.pdf. – Cité page 68.

[Zweigenbaum et al., 2001] Zweigenbaum, P., Jacquemart, P., Grabar, N., and Habert, B. (2001). Building a text corpus for representing the variety of medical language. *Stud Health Technol Inform*, 84(Pt 1):290–294. – Cité pages 17 et 85.

[Zweigenbaum et al., 2016] Zweigenbaum, P., Moriceau, V., and Grouin, C. (2016). LIMSI at CEGS N-GRID 2016 NLP Shared-Tasks: track 2, RDoC. In *Proceedings of the CEGS N-GRID Work*, Chicago, IL. – Cité page 49.

[Zweigenbaum et al., 2012] Zweigenbaum, P., Wisniewski, G., Dinarelli, M., Grouin, C., and Rosset, S. (2012). Résolution des coréférences dans des comptes rendus cliniques. Une expérimentation issue du défi i2b2/VA 2011. In *Actes de RFIA*, Lyon, France. hal-00656514. – Cité page 49.

Chapitre

Table des figures

Chapitre

Liste des tableaux

Index

Abstract

In this habilitation thesis, I present the research I have conducted on language productions from speakers made on social networks. This habilitation thesis is organized into two main parts: first, the impact of speakers on their language, the latter being considered here as the object of study, and second, the impact of users on the tools and resources used in natural language processing. Due to cultural differences on the one hand, and technical and societal evolutions on the other hand (such as the use of French inclusive language), the NLP domain is in constant evolution in order to tackle this linguistic variability, which is representative of the individual diversity. We have considered the opportunity to study inferences for opinion mining in Chinese, as a complementary way to the identification of emotion/sentiment/opinion words. Social networks are a valid source of relevant testimonies in pharmacovigilance, for the detection of side effects or drug misuse, and in a pandemic context. While computer technology now makes it possible to encode more information, especially statistical information, and although gender stereotypes have been identified in current transformers models, our work combining morpho-syntactic information with vector representations confirms the complementary of linguistic information in several classical NLP tasks. The future scientific obstacles to be removed will come from the now more present imbrication of multimodality in language productions.

Résumé

Dans ce manuscrit, je présente les recherches que j'ai menées sur les productions langagières des locuteurs d'une langue sur les réseaux sociaux. Mon manuscrit s'articule autour de deux angles d'analyse : l'impact des utilisateurs sur la langue, cette dernière étant alors envisagée comme objet d'étude, et l'impact des utilisateurs, au travers de leurs productions langagières, sur les outils et ressources utilisés pour le traitement automatique des langues. Face aux différences culturelles d'une part et aux évolutions techniques et sociétales d'autre part, telle que l'apparition du français inclusif, le traitement automatique des langues est lui-même en constante évolution pour faire face à cette variabilité linguistique, représentative de la diversité individuelle. Nous avons constaté l'opportunité d'étudier les inférences pour de la fouille d'opinion en chinois en complément des mots porteurs d'opinion/sentiment/émotion. Les réseaux sociaux constituent une source de témoignages pertinente en pharmacovigilance pour la détection des effets secondaires ou du mésusage médicamenteux, ou encore en contexte pandémique. Alors que l'informatique permet désormais d'encoder davantage d'informations, notamment d'ordre statistique, et bien que des stéréotypes de genre aient été identifiés dans les modèles transformers actuels, les travaux combinant des informations morpho-syntaxiques aux représentations vectorielles confirment la complémentarité des informations linguistiques dans plusieurs tâches classiques du TAL. Les prochains verrous scientifiques à lever viendront de l'imbrication désormais plus marquée de la multimodalité dans les productions langagières.

Date du document : 27 janvier 2023