Université d'Orléans
UFR Lettres Langues et Sciences Humaines

Mémoire de fin d'études

Spécialité : *Linguistique outillée et traitement automatique des langues*

Thème

Documentation des pratiques prescriptives chez les non-experts de la langue française sur les réseaux sociaux.

Encadré par
Emmanuel Schang

Réalisé par
Maël Cheval

Lien GitHub : https ://github.com/Mael45000/memoireM2/tree/main

2024

0.1 Remerciements

Tout d'abord, merci à M.Schang pour m'avoir aiguillé sur mon choix de problématique.

Je remercie aussi Anne Abeillé et Heather Burnett de m'avoir accompagné et d'avoir répondu à toutes mes interrogations pendant le stage. Ma gratitude se porte aussi sur Marie Flesch, pour m'avoir aiguillé sur mes choix de méthodologie.

Enfin, j'aimerais remercier mes proches, pour leurs soutiens ainsi que leurs précieux conseils pendant toute cette période.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

La tradition prescriptive en France est particulièrement importante et nous offre un éventail d'ouvrages, dont la création d'un dictionnaire par l'Académie française (AF-, 1694) au XVIe siècle. Les pratiques prescriptives réfèrent à l'ensemble des comportements qui renvoient au respect de la norme dans le langage. Par exemple, il nous ai tous déjà arrivé de corriger un proche qui fait une faute lors d'une conversation (*On ne dit pas « aller au coiffeur » mais « aller chez le coiffeur » !*). La documentation des pratiques prescriptives chez les experts de la langue, c'est-à-dire les linguistes, les auteurs d'ouvrages ou les enseignants est déjà assez présente. Cela paraît assez logique, étant donné que ce sont les experts de la langue qui sont amenés à parler de celle-ci. On peut citer les blogs spécialisés sur les questions de langue, par exemple.

Le métadiscours chez les non-experts, c'est-à-dire les individus qui ne portent pas d'intérêt particulier à la langue, est une pratique courante. Avec l'arrivée d'Internet, les échanges entre individus se sont multipliés, tout comme les pratiques prescriptives. Cependant, on ne retrouve que très peu de documentation de ces pratiques chez les non-experts. Cela s'explique par le fait qu'il est difficile de documenter ces pratiques en ligne, au contraire des experts. Afin de capturer ces pratiques dans un environnement naturel, les réseaux sociaux semblent être une bonne solution. Osthus (2018) avait déjà relevé la nécessité de recueillir les sources des « locuteurs ordinaires » en dehors des espaces de discussion explicitement portés sur le métadiscours, afin de mieux comprendre le fonctionnement

du métalangage au quotidien. Le métalangage est le fait de parler de la langue et de son fonctionnement.

Pour documenter un sujet, il est possible de prendre des corpus déjà existants afin d'extraire les informations que l'on souhaite. Il est aussi envisageable de créer son propre corpus afin d'avoir précisément les informations que l'on cherche, c'est le cas quand il n'y a pas encore d'études réalisées sur un sujet. Pour les deux options, ces tâches demandent l'intervention d'un outil de traitement automatique des langues, que ce soit pour explorer un corpus ou en créer un.

Dans le cadre d'une recherche de pratiques prescriptives chez des non-experts de la langue, on peut se questionner sur les types de « fautes » qui seront relevées par ces internautes. Le problème principal est de savoir comment relever ses pratiques dans l'océan d'interactions des réseaux sociaux. **Comment peut-on identifier et extraire les pratiques prescriptives des non-experts de la langue sur les réseaux sociaux ?** Doit-on uniquement se concentrer sur quelques exemples populaires de fautes en français ou peut-on espérer récolter un plus large panel de pratiques prescriptives ? Une solution serait de se pencher vers des techniques utilisant des outils de traitement automatique des langues, qui peuvent permettre l'identification, l'extraction et l'exploration de ces pratiques langagières. On émettra l'hypothèse qu'on pourra difficilement extraire une majorité des pratiques prescriptives et qu'il faudra se concentrer sur des exemples connus de fautes en français. Nous supposerons aussi que les réseaux sociaux sont un lieux propice à l'apparition de notre phénomène.

Ce mémoire est divisé en trois parties. La première consiste à définir le cadre de travail, la seconde décrit les processus d'extractions des termes et le dernier porte sur une analyse des résultats des extractions. Nous pourrons ensuite répondre à notre problématique dans la conclusion. Tous les documents qui seront cités dans ce mémoire se trouvent sur un projet GitHub [1] spécialement crée pour cela.

1. https ://github.com/Mael45000/memoireM2/tree/main

Chapitre 2

Présentation du cadre de travail et définition de l'étude

2.1 Environnement de travail

L'organisme d'accueil de mon stage est le laboratoire de Linguistique Formelle (LLF). Situé au sein de l'Université Paris Cité, ce laboratoire traite le langage sous tous ses angles. Les thèmes de recherche s'axent sur la grammaire du mot, la grammaire de la phrase et la grammaire de l'énoncé et du discours. Il a été fondé en 1972 par Antoine Culioli et était à la base une équipe de recherche associée.

Le laboratoire est membre du Labex Empirical Foundations of Linguistics. C'est un laboratoire qui rassemble des équipes de linguistique dans le but d'ouvrir de nouvelles perspectives en partageant les données et les réflexions. Ce laboratoire permet d'avoir de nombreuses collaborations. La spécificité que revendique le LLF est de compter une grande diversité ; qu'elle soit dans les approches théoriques, les méthodes empiriques ou encore dans les langues étudiées.

Parmi les ouvrages scientifiques récents des membres du laboratoire, on peut citer De Clercq *et al.* (2023) avec « Adverbial resumption in verb second languages »,Donati et Graffi (2022) avec « La grammatica generativa » ou encore

Abbou (2022) avec « Tenir sa langue ».

2.2 Modalités du stage

2.2.1 Contexte du stage

Le stage se contextualise dans le projet ANR Mathy. Ce projet a pour objectif de développer un nouvel axe d'étude, entre linguistique et philosophie des mathématiques, se nommant « mathematical hygiene » qui est définie comme étant « l'ensemble des discours normatifs régulant les pratiques mathématiques. » Arana et Burnett (2023). Ce terme est inspiré de celui de Cameron (1995) avec l'expression « verbal hygiene », qui définit les phénomènes normatifs que les locuteurs d'une langue peuvent exprimer à propos de certaines pratiques linguistiques.

Le projet se découpe en trois principaux objectifs. Le premier est de documenter des pratiques d'hygiène mathématiques. Le second est de comparer ces pratiques avec des phénomènes d'hygiène verbale. Le dernier objectif est de comprendre comment les phénomènes d'hygiène verbale et mathématique agissent dans la psychologie du raisonnement.

Mon stage se déroule dans le cadre du deuxième objectif du projet ANR Mathy. Plus précisément, l'objectif de mon stage est de documenter les pratiques d'hygiène verbale dans des médias traditionnels et dans les réseaux sociaux. La recherche se tourne en particulier dans les discours féministes, car ils sont exposés à beaucoup de commentaires normatifs, mais aussi parce qu'ils sont de plus en plus nombreux sur les réseaux sociaux ces dernières années.

2.2.2 Travail effectué

Ma principale mission lors de ce stage était d'annoter un jeu de données axé sur le pronom iel [1]. Ce pronom étant assez récent, son entrée dans le dictionnaire

1. Le Robert définit le pronom iel comme « un pronom personnel sujet de la troisième personne du singulier (iel) et du pluriel (iels), employé pour évoquer une personne quel que

a provoqué de nombreuses réactions sur les réseaux sociaux. Les données avec lesquelles j'ai travaillé proviennent de trois réseaux sociaux : Reddit, X (Twitter) et Youtube.

Le travail d'extraction des données étant déjà réalisé, ma mission portait sur le tri et l'annotation des données déjà structurées. Le tri des données permet de supprimer les interventions qui sont des cas d'usage du pronom iel, et de garder ceux qui sont des commentaires. Les cas d'usage correspondent à une utilisation standard du pronom iel. Les commentaires font référence aux interventions métalinguistiques sur le pronom iel. Ce sont ces commentaires qui nous intéressent puisqu'ils sont susceptibles de contenir des pratiques prescriptives.

Ensuite, à l'aide de mes tutrices de stage, Anne Abeillé et Heather Burnett, nous avons élaboré un guide d'annotation pour chaque catégorie d'annotation présente dans le tableau 3.1 ci-dessous. Ces catégories d'annotation ont été créées dans le but d'avoir une description précise des pratiques prescriptives. Cela signifie que ces catégories n'étaient pas prédéfinies et que certaines d'entre elles ont été ajoutées entre temps. Les catégories « Critique utilisateur » et « Langage/genre » ont été ajoutés plus tard, car ces attributs ne pouvaient pas rentrer dans les catégories déjà présentes.

Nous avons dû faire face à des difficultés qui limitent la qualité de l'annotation. Premièrement, les données ont été extraites en prenant un empan limité de mots autour du pronom « iel ». Le problème est que certains commentaires sont trop longs et ne prennent pas tout le contexte, ce qui nous empêche parfois de comprendre le commentaire. Le second élément qui peut poser problème est le cas du sarasme, car il peut être complexe à cerner. Le risque est de ne pas comprendre l'intention principale du commentaire.

soit son genre. »

Intitulé	Définition	Exemple	Annotation
Domaine	Le sujet sur lequel le commentaire porte.	C ' est moche iel , il y aurait fallut inventé autre chose ...	esthetique
Polarité	Le commentaire se présente comme étant positif, négatif ou neutre vis à vis du pronom iel.	Alors iels et celleux sont quand même vachement pratiques , parfois !	positif
Prescriptif	Le commentaire a-t-il une tendance prescriptive pour le pronom iel ?	Absolument contre le " iel "!!!	oui
Critique utilisateur	L'utilisateur critique les utilisateurs de iel.	Le jour où tu utilises iel dans un mail de manière non ironique , mp stp	oui
Langage/ genre	Le commentaire peut porter sur le l'aspect linguistique de iel (langage) ou sur la personne qui l'utilise (genre).	@JB1180Tv Pourquoi tu dis iel ? C ' est Lola , donc , très probablement un « elle » , il faut arrêter de supprimer le genre	genre
Emotion	Emotion ressentie dans le commentaire.	@Ascane09 anti iel le user, la honte un peu, j'oserai même pas afficher mes idées dans un @	honte

TABLE 2.1 – Catégories d'annotation du pronom iel.

2.3 Définitions

Cette partie est consacrée aux définitions primordiales à la bonne compréhension de la problématique et du travail effectué par la suite.

2.3.1 Expert et non-expert

Au cours de ce mémoire, les termes d'experts et de non-experts seront utilisés pour parler des individus qui sont spécialistes de la langue et ceux qui ne le sont pas.

Selon Osthus (2018) chaque locuteur est en quelque sorte un expert de la langue, avec tout de même une certaine hiérarchie d'expertise. Par exemple, un linguiste sera considéré comme étant plus expert qu'un locuteur ordinaire. L'auteur souligne la difficulté de donner une définition des personnes non-expertes de la langue :

> La catégorie la plus difficile à définir est celle des « locuteurs ordinaires ». Dans le cadre de notre analyse, on ne peut fournir qu'une définition négative. C'est quelqu'un qui n'est ni spécialiste des sciences du langage, ni amateur de la langue (par exemple particulièrement enthousiasmé par les questions de défense de la langue).

Nous nous tiendrons à cette définition de non-expert, car elle permet de délimiter l'expert d'un non-expert. La distinction est d'autant plus difficile que les locuteurs sur Internet sont majoritairement anonymes. Si les locuteurs ne communiquent pas via une plateforme explicitement linguistique (par exemple, un billet de blog sur un sujet de langue), il est quasiment impossible de connaître leur catégorie d'expertise. Il est possible de retrouver des personnes expertes sur les réseaux sociaux, que ce soit explicite (la personne qui parle n'est pas anonyme et est experte de la langue) ou implicite (la personne se revendique experte, bien qu'elle soit anonyme). Afin de faciliter la catégorisation des internautes, si ce dernier provient d'un réseau social, il est considéré comme étant non-expert de la langue. Il faut donc prendre en compte qu'il y a une marge d'erreur dans la catégorisation des experts et des non-experts.

2.3.2 Prescriptisme et descriptisme

L'utilisation de la langue est soumise à deux points de vue antagoniste. La démarche prescriptive est opposée à la démarche descriptive en linguistique.

Le prescriptivisme est, selon Hare et Mervyn (1952) (citée dans Dutant (2012)), « une doctrine méta-éthique qui voit dans les jugements moraux des prescriptions que le sujet fait à lui-même et à autrui ». Le prescriptivisme se base sur une autorité dans la langue qu'il faut suivre, dans le but de parler parfaitement la langue. En France, c'est l'Académie française [2] qui fait figure d'autorité.

Au contraire, le descriptivisme tend à penser que l'usage de la langue ne doit pas être limité aux normes présentes, mais qu'on doit laisser la langue évoluer en acceptant les changements.Andrews (2006) défini la grammaire descriptive comme « une approche qui décrit la langue telle qu'elle est utilisée ».

2.3.3 Faute linguistique

Selon le trésor de la langue française informatisé (TLFI), une faute [3] est « un manquement à une règle morale, à une règle de conduite, une action considérée comme mauvaise ». Si l'on rapporte cette définition dans le domaine de la linguistique, on considère qu'une faute est l'ensemble des comportements linguistiques se trouvant en dehors de la norme d'une langue donnée.

Un comportement prescriptif visera à faire appliquer la norme, ce qui implique de sanctionner les fautes. Un comportement descriptif va prendre en compte la faute, mais ne va pas la sanctionner.

2.4 Etat de l'art

Cet état de l'art est d'abord consacré au sujet de recherche : les pratiques prescriptives. Nous allons ensuite faire un tour d'horizon des travaux sur la

2. https ://www.academie-francaise.fr/
3. https ://www.cnrtl.fr/lexicographie/faute

méthode nous permettant de répondre à la problématique : l'extraction d'information.

2.4.1 Documentation de pratiques prescriptives

Les études sur la documentation de pratiques prescriptives chez les individus non-experts sont peu nombreuses. Cependant, si l'on aborde un regard plus large, on retrouve des études sur les pratiques prescriptives sur Internet.

Tavosanis (2007) a réalisé une typographie des erreurs d'orthographe sur le Web. Dans celle-ci, les erreurs peuvent être intentionnelles et dues à un manque de connaissances de la langue de la part du locuteur. Cela peut être le cas si on ne connaît pas l'orthographe d'un mot. Les fautes sont parfois intentionnelles. C'est le cas par exemple quand il est impossible d'écrire un mot en raison de l'encodage qui empêche d'écrire certains caractères. Dans cette étude, l'auteur ne mentionne pas si les locuteurs sont considérés comme experts ou non-experts.

Damar (2010) va examiner des discours normatifs sur Internet. Son objectif est d'observer s'il est possible de catégoriser les expressions du purisme, en fonction de variables telles que le lexique ou la syntaxe. Seuls trois faits de langue ont été choisis pour documenter le purisme linguistique ; une expression (aller au coiffeur) et deux types de néologismes (ceux dus aux mots féminisés et ceux dus aux nouvelles technologies). Il est intéressant de noter que l'auteur a établi un corpus en récoltant des discours de forums n'étant pas forcément spécialisés dans la langue française. Cela implique qu'on peut retrouver dans ce corpus des productions de purisme linguistique chez des experts et des non-experts de la langue. Cependant, dans cet article, il n'y a pas de classification réalisée entre les individus.

Amadieu (2014) va rechercher les irrégularités linguistiques des auteurs classiques selon les manuels de rhétorique française au XIXe siècle. Ces manuels vont avoir une autorité sur la norme linguistique de l'époque. Ici, la documentation est consacrée à des productions d'experts de la langue.

L'apparition du langage SMS a toujours fait débat, car il est éloigné de la

norme. Moïse (2015) à étudié le regard des locuteurs sur l'écriture SMS afin d'observer un éventuel discours normatif sur celui-ci. Il n'est donc pas question de critiquer l'emploi d'un terme qui ne serait pas dans la norme, mais l'idée globale de l'écriture SMS. Les locuteurs critiquent en majorité l'aspect inquiétant de s'éloigner de la norme. L'auteur ne discerne pas les individus selon leur expertise, mais les métadonnées sur leur catégorie sociale, leur âge ou leur orthographe peut donner des indices sur leur expertise.

Par le biais d'une étude sur des articles de journaux en ligne, Calabrese et Rosier (2015) s'intéressent au discours normatif des lecteurs de journaux en ligne. Leur documentation comporte une trentaine d'articles provenant de journaux en ligne, mais aussi des billets de blog et des exemples sur le réseau social Facebook. Les auteurs vont qualifier les discours des lecteurs de « non-professionnels » et les discours des journalistes de « professionnels » ou d' « expert ». Dans ce cas, nous avons une récolte de commentaires métalinguistiques d'individus qui ne sont pas considérés comme des experts de la langue par les auteurs. Cet article ne précise pas la méthodologie appliquée pour récupérer ces commentaires.

Plus récemment, Humphries (2023) a réalisé une étude sur les comportements prescriptifs des internautes sur la plateforme X (anciennement Twitter). L'auteur a décidé de s'intéresser aux fautes « croyent » et « voyent » pour constituer son corpus. C'est donc un corpus uniquement constitué de pratiques prescriptives de non-experts de la langue.

En conclusion, on peut dire que les travaux concernant la documentation de pratiques prescriptives sur Internet portent sur des sujets variés. La séparation entre expert et non-expert n'est pas toujours présente, car elle n'est simplement pas le sujet des travaux, même si on retrouve une catégorisation des locuteurs chez Damar (2010) par exemple. Le travail de Humphries (2023) est celui qui est le plus proche de notre objectif de documentation de pratiques prescriptives de non-experts, même si la méthodologie choisie est assez simpliste, car elle cible deux termes dans un corpus.

Nous allons maintenant regarder les avancées des travaux dans le domaine de l'extraction d'information.

2.4.2 Extraction d'information

L'extraction d'information fait partie des sous-domaines du traitement automatique des langues (TAL), comme peut l'être le traitement du signal [4] par exemple. L'extraction d'information comprend aussi une multitude de domaines. On peut citer l'analyse de sentiments, qui permet d'extraire le ressenti d'un texte grâce aux mots présents. La fouille de textes est le sous-domaine qui nous permettra de répondre à notre problématique, car il permet de rechercher des informations dans un corpus de données textuelles.

L'idée de vouloir trouver des informations rapidement et de manière précise est née dés l'apparition des premiers ordinateurs. La volonté des chercheurs était d'abord d'automatiser la recherche des livres dans les bibiliothèques. Bush (1945), va être le premier à conceptualiser les liens hypertextes. [5] La méthode de recherche utilisée était une recherche booléenne ; c'est à dire une variable comprenant deux états, qui représentent les valeurs de vérité « vrai » ou « faux ». Elle sert à vérifier la présence ou l'absence d'un mot-clé lors d'une requête.

La méthode booléenne étant limitée, notamment dans la prise en compte de la sémantique, de nouvelles méthodes vont voir le jour. Le modèle vectoriel va être introduit par Salton *et al.* (1975) et permet de considérer la dimension sémantique d'un document. Elle va donner la représentation mathématique d'un texte.

A la même période, Robertson et Jones (1976) proposent un modèle probabiliste de pertinence. Ce dernier a pour objectif de mesurer la probabilité de la pertinence d'un document selon une requête prédéfinie.

L'arrivée d'Internet va entraîner une explosion des données non structurées, c'est-à-dire des données qui ne sont pas stockées dans un format prédéfini. Ce contexte va pousser les chercheurs à trouver de nouvelles solutions plus effi-

4. Le traitement du signal est un sous domaine du TAL. Ce domaine permet de traiter du signal sonore ou graphique de manière automatique. Il comprend lui-même des sous-domaines, comme la reconnaissance automatique de la parole ou la reconnaissance automatique de caractères.

5. D'après le Larousse, l'hypertexte est une "technique ou système permettant, dans une base documentaire de textes, de passer d'un document à un autre selon des chemins préétablis ou élaborés lors de la consultation".

caces pour distinguer les informations linguistiques. Ainsi, Grishman (1997), va chercher à identifier les coréférences [6] et les inférences [7] dans ces données non structurées.

De nos jours, le domaine de l'intelligence artificielle a permis une avancée dans l'extraction d'information, à l'aide de l'apprentissage automatique. Ce champ d'étude fonctionne grâce à des approches statistiques. Le but est d'entraîner un algorithme à résoudre une tâche. Les réseaux de neurones récurrents sont un type de réseaux de neurones artificiels. Ils sont constitués d'unités connectées entre elles. Les calculs de ses neurones se font en prenant en compte l'information à une période antérieure (Pour l'instant t, on prend les informations de t-1.). Le problème principal de ce modèle est qu'il n'arrive pas à mémoriser les actions passées. C'est ce que le modèle Long short term memory (LSTM) de Hochreiter et Schmidhuber (1997) va parvenir à résoudre.

Ce sont actuellement les grands modèles de langage qui sont les plus utilisés dans la majorité des domaines, car ils sont performants et efficaces dans un grand nombre de tâches. Les grands modèles de langage comme BERT (Delvin *et al.*, 2018), permis notamment par l'architecture Transformer [8] (Vaswani *et al.*, 2017) comptent des milliards de paramètres qui permettent de résoudre ces tâches. Ce type de modèle possède cependant quelques limites. En effet, ces modèles de langage reflètent les données avec lesquelles ils sont entraînés. Cela inclut qu'il ne peut pas couvrir un sujet dans son ensemble. Si le modèle est entraîné sur des données françaises, il donnera des réponses culturellement centrées sur l'Europe. Ensuite, ces modèles fonctionnent en générant une réponse qui est statistiquement la plus attendue dans le contexte, mais il peut arriver que le système se trompe et donne une réponse fausse en affirmant sa source. Enfin, ces modèles sont entraînés sur des quantités énormes de données, qui

6. D'après le Larousse, une coréférence est « une relation entre deux termes ayant le même référent. » Par exemple, dans les phrases « Paul mange son déjeuner. Il l'adore. » , « Paul » et « Il » renvoient à la même personne. Ce sont donc des coréférents.

7. D'après le TLFI, une inférence est « une opération qui consiste à admettre une proposition en raison de son lien avec une proposition préalable tenue pour vrai ». Dans l'exemple « Il y a une glace au chocolat et une glace à la fraise. Jules n'aime pas le chocolat. » Jules choisira la glace à la fraise, car la proposition préalable tenue pour vrai est qu'il n'aime pas le chocolat.

8. Contrairement aux réseaux de neurones récurrents, cette architecture va analyser une phrase en entrée en portant attention à plusieurs parties de la phrase en même temps. Cela permet d'entraîner des modèles plus rapidement que les précédents modèles.

sont parfois protégées. La génération de texte étant inspiré de ces données, cela pose un problème éthique, même si aucune loi ne sanctionne pour l'instant ce comportement.

En résumé, nous cherchons à extraire des informations depuis que nous avons eu la possibilité de stocker des données de manière numérique. L'évolution de la masse de données au cours du temps nous a contraints à adapter les techniques d'extraction de l'information, jusqu'à l'arrivée d'Internet et une explosion de la quantité d'informations. Les nouvelles techniques d'apprentissage automatique sont maintenant favorisées pour leur efficacité et leur polyvalence.

Chapitre 3

Méthodologie et extraction des pratiques prescriptives

Ce chapitre est la pièce centrale du mémoire. Nous verrons comment nous allons procéder à l'extraction des pratiques prescriptives chez des non-experts de la langue. Pour cela, nous verrons le corpus utilisé, l'outil choisi et bien sûr la méthodologie pour extraire les données.

3.1 Choix du corpus

Dans cette partie, nous allons voir à partir de quelles données nous allons appliquer l'extraction des pratiques prescriptives. C'est un choix crucial, car elles peuvent influencer sur le résultat de la méthode. Les données se doivent d'être pertinentes avec le thème de recherche.

3.1.1 Paramètres de la séléction

Tout d'abord, nous devons nous diriger vers des données qui répondent aux exigences de la problématique. La première est de trouver des pratiques prescriptives. Ce type de données peut se retrouver dans beaucoup de situations d'interactions, ce qui laisse beaucoup de choix dans le type de données à choisir. La seconde exigence que nous avons est de cibler des locuteurs non-experts de la langue. Il suffit d'éviter les communautés et les institutions se revendiquant

expertes de la langue. Sur Internet, les endroits où l'on trouve le plus de pratiques prescriptives chez des locuteurs non-expert sont les réseaux sociaux.

Dorénavant, il faut savoir comment accéder aux données. Il est possible de prendre un corpus déjà existant qui possède des données provenant des réseaux sociaux, mais nous pouvons aussi créer notre propre corpus en choisissant d'extraire directement les commentaires. Le scrapping » est le terme utilisé pour extraire des données sur Internet. En raison du droit de la propriété intellectuelle et commerciale, il n'est pas possible de scrapper n'importe quelles données, au risque d'être sanctionné juridiquement. Il existe des outils qui permettent l'extraction de ces données sans codage comme Octoparse [1] ou Apify [2]. Par exemple, l'API [3] Reddit Scrapper [4] est un outil crée dans le but d'extraire toutes les métadonnées du réseau social Reddit. Cependant, ce type d'outil est généralement payant ou propose des choix limités dans le paramétrage des recherches. La requête se limite à un mot-clé sur le site, ce qui n'est pas suffisant pour établir un corpus de pratiques prescriptives.

L'alternative à la création d'un corpus spécialisé dans les pratiques prescriptives est d'utiliser un corpus déjà existant et d'y extraire les données. Nous ciblons la recherche sur des corpus comprenant des données sur les réseaux sociaux. Afin d'optimiser les chances d'avoir des pratiques prescriptives, nous privilégions les corpus avec une grande quantité de données. Le site Ortholang [5] est une plateforme qui rassemble plus de 700 ressources linguistiques dans le cadre du traitement de la langue française. On y retrouve des outils, des terminologies, des lexiques, mais aussi des corpus. Le corpus choisi se nomme « TREMoLo-Tweets ».

1. https ://www.octoparse.fr/
2. https ://apify.com/
3. D'après la Commission nationale de l'informatique et des libertés (CNIL), une API (application programming interface ou « interface de programmation d'application ») est une interface logiciel qui permet de « connecter » un logiciel ou un service à un autre logiciel ou service afin d'échanger des données et des fonctionnalités.
4. https ://apify.com/trudax/reddit-scraper
5. https ://www.ortolang.fr/fr/accueil/

3.1.2 Présentation du corpus

Le corpus « TREMoLo-Tweets » (Mekki, 2023) a été créé dans le cadre du projet « TREMoLo ». [6] Son objectif est de classifier les emplois des registres de langue, puis d'automatiser la classification de ces emplois. Comme son nom l'indique, ce corpus est composé de tweets français (on compte 228 505 tweets, soit un total de 6 millions de mots) L'automatisation des annotations se fait avec le modèle de langage CamemBERT [7] (Martin *et al.*, 2020), qui est un modèle entrâiné sur RoBERTa [8] (Liu *et al.*, 2019), lui-même entrâiné sur BERT.

Le corpus se structure sous la forme d'un fichier de type CSV de 228 505 lignes et 6 colonnes. Chaque ligne représente un tweet, les 6 colonnes sont respectivement :

— *Text_ index* qui représente l'identifiant du tweet.
— *Familier, Courant, Soutenu et Poubelle* qui représentent la proportion des registres de langue qui a été donné par le classifieur.
— *Texte* qui illustre le texte du tweet.

6. https ://anr.fr/Projet-ANR-16-CE23-0019
7. CamemBERT est un modèle de langage francais basé sur l'architecture du modèle de langage RoBERTa. Il permet d'évaluer l'étiquetage des parties du discours, la reconnsaissance d'entité nommées ou encore les inférences en langage naturel.
8. RoBERTa est un modèle de langage qui reprend l'architecture BERT dans le but d'améliorer les performances du modèle. L'idée est de pré entraîner le modèle en prêtant attention aux choix des hyperparamètres.

	Text_index	Familier	Courant	Soutenu	Poubelle	Texte
3	82013_tweet.txt	0.9960266351	0.0107797086	0.0002260807	0.0001506505	Bosh il lui a fait un sal boulot à kaaris il lui a montré que maintenant y'a plus de petit et grand
5	130701_tweet.txt	0.9961282014	0.0114650726	0.0002101367	0.0001553595	Le craquage est complet, j'ai préco le repackage de straykids, et Mario Allstar 😊🤡😊
7	39077_tweet.txt	0.9987241029	0.0032797157	0.0005043147	0.0004163087	Allez c'est parti #Cdiscount2emeDemarque
18	98773_tweet.txt	0.9986186027	0.0035552976	0.0004616677	0.0003715757	@XXX @XXX Chut faut pas le dire à Camille
19	258184_tweet.txt	0.9978873729	0.0059151351	0.0003098847	0.0002419657	ma passion pendant le générique des vlogs d'août: essayer de deviner qui est qui quand y'a les dessins des potes à Léna
20	262936_tweet.txt	0.9957979917	0.0105948746	0.0002384187	0.0001452567	Me encontré #MTVHottest BTS @XXX url_path
21	99082_tweet.txt	0.9989230632	0.0017877221	0.0009896757	0.0005716667	@XXX putain, ils veulent notre peau Nintendo. C'est day one les amiibo
23	64005_tweet.txt	0.9975546002	0.0070988833	0.0002756417	0.0002152027	RT @XXX: Je suis en larmes quel roi 😭😭 url_path
30	90158_tweet.txt	0.9984832406	0.0041000247	0.0004091857	0.0003357527	Je suis POUR les grèves mais alors encore une de la #SNCF sachant qu'on est en plein Covid et qu'on s'entasse déjà en temps normal, je v
34	189761_tweet.txt	0.9678494930	0.0619827806	0.0001665657	5.948269972	Maroua a chaque drama elle me déçoit un peu plus je suis choquée d'elle
48	92675_tweet.txt	0.9988892078	0.0015259386	0.0011973977	0.0005752447	@XXX @XXX @XXX Des grand Denjiro malina Marié, tout sa pour procréer un énergumènes tel que jctrrrr eeeesh foorce a vous deux hein on
49	71292_tweet.txt	0.9988860487	0.0015254020	0.0011989177	0.0005714557	Même si on lui coupe le pied il doit jouer. Back to back MVP, MVP et DPOY sur la même saison et tu vas te manger un sweep ????? LMAO
51	160768_tweet.txt	0.9534457921	0.0832462015	0.0001717807	5.433131445	#Hassan les personnes qui croient encore que Hassan a frapper fathma venez on se pose 2 min et on discute
53	195174_tweet.txt	0.9988853931	0.0014248490	0.0013180977	0.0005923507	j'espère grave que Léna elle va passer un moins entier chill sans emmerde (et sans maladie) parce que bichette toujours elle a la poisse
57	35968_tweet.txt	0.9973284006	0.0075787901	0.0002614257	0.0001983647	@XXX Mais non mais imagine elle me recale ...
65	148387_tweet.txt	0.9989142417	0.0015810132	0.0011518007	0.0005878217	Le pauvre Westbrook mais oklm il fais un bon match url_path
71	58099_tweet.txt	0.9989072680	0.0017950832	0.0009784957	0.0005540547	Sa soule trop de changement sur #rmclive #RMCF après bourdin,brunet maintenant fatima j aime pas tous ses changements vous aller nous
73	233171_tweet.txt	0.9942213892	0.0164034366	0.0001832547	0.0001210147	@XXX En contre, mais on les négocie mal, et en défense ben chaque fois que Brest arrive dans notre moitié, ils peuvent marquer mdr.
74	188152_tweet.txt	0.9989079236	0.0015522535	0.0011812447	0.0005896097	jamais tranquille c un truc de malade même contre brest
79	243462_tweet.txt	0.9988796710	0.0013576746	0.0014143887	0.0006046897	J'aurais parlé comme ca ma mere m aurais mis un coup de tong dans la gueule voir un coup de poele mdr #LRDS
80	1723_tweet.txt	0.9988777637	0.0023242235	0.0007252957	0.0005050007	le pauvre Jean-Luc Gary il tremble a mort #SNCF
81	97010_tweet.txt	0.9989098906	0.0017661154	0.0010151567	0.0005814437	On en parle du bop de Somi ??
87	88325_tweet.txt	0.9959012866	0.0114737451	0.0002206067	0.0001507997	@XXX @XXX Vas rejoindre David Luiz qu'on t'insulte tes grands morts @XXX
94	45323_tweet.txt	0.9988392591	0.0026647746	0.0006229877	0.0004759727	@XXX Qu'est ce que jferais pas pour toi dis donc
103	203785_tweet.txt	0.9988681077	0.0017697215	0.0010062757	0.0005466937	Bh clairement personne n'aime les poissons jpp, j'ai le pire des signes astro 😭
104	42353_tweet.txt	0.9989132881	0.0015583036	0.0011816627	0.0005947057	Kunigami il s'est fait jeté j'ai le seum url_path

FIGURE 3.1 – Extrait du corpus TREMoLo-Tweets

Cette partie nous a permis de cibler le type de corpus adapté à la recherche de pratiques prescriptives de non-expert. Nous avons fait le choix d'un corpus contenant un large panel de tweet dans l'objectif de trouver notre phénomène. La discussion se tourne maintenant sur le procédé technique de manipulation du corpus.

3.2 Choix de l'outil

3.2.1 Utilisation d'un logiciel de TAL ?

Il est important de bien choisir l'outil qui va nous permettre de manipuler les données du corpus. Le choix peut se porter vers des outils spécialement consacrés au traitement automatique du langage, comme par exemple TXM [9]. L'outil TXM tient son nom du domaine auquel il traite ; la textométrie. Ce dernier est né dans les années 1970 en France. Il consiste en l'analyse de données linguistiques à l'aide de modèles statistiques, ce qui permet d'avoir un traitement plus quantitatif de ces données. La textométrie se rapproche de l'extraction d'information, car ils ont tous deux la volonté de traiter des données textuelles en

9. https ://txm.gitpages.huma-num.fr/textometrie/

grande quantité. Toutefois nous pouvons les différencier, car l'extraction d'information vise à récuperer des données textuelles via des corpus, alors que la textométrie va explorer un corpus dans le but d'en faire ressortir ses caractéristiques linguistiques.

L'outil TXM a l'avantage d'être simple d'utilisation pour ceux qui utilisent rarement des logiciels, mais est aussi d'utiliser le langage de codage R pour les initiés. L'outil permet de faire des requêtes de mots, mais aussi de lemmes [10] grâce à l'étiqueteur morphosyntaxique TreeTagger [11]. Un étiqueteur morphosyntaxique permet d'annoter automatiquement les informations grammaticales associées à un mot. TXM permet également de créer des sous-corpus qui peuvent ensuite être convertis en tableaux. Cette fonctionnalité est pratique si l'on veut par exemple récolter des occurrences de termes spécifiques. Cet outil est donc intéressant pour explorer un corpus et y extraire des données.

Le logiciel se bute cependant à des limites pour extraire correctement des pratiques prescriptives. Premièrement, il est nécessaire de procéder à un prétraitement des données du corpus, ce que TXM ne peut pas réaliser. Deuxièmement, la recherche de pratiques prescriptives ne se limite pas à l'utilisation de requêtes. Si la méthode consiste à détecter des fautes d'orthographe, alors il n'est pas pertinent d'utiliser TXM, car il n'a pas de fonctionnalité qui permette ce genre d'action. En résumé, TXM est un logiciel complet dans le cas de l'analyse d'un ou des plusieurs termes spécifiques. En ce qui concerne notre phénomène, il est moins pertinent de l'utiliser.

3.2.2 Utilisation d'un langage informatique ?

3.2.2.1 Présentation de Python

L'utilisation d'un logiciel spécialisé en traitement automatique des langues comme TXM ne nous permet pas d'envisager l'ensemble des méthodes de récolte de notre phénomène. La solution serait d'utiliser un langage de programmation qui nous permettrait d'avoir la mainmise sur l'ensemble des méthodes envisagées. Dans le domaine du TAL, le langage de programmation le plus utilisé est

10. Ce terme est utilisé pour parler de la forme du mot utilisée dans l'entrée des dictionnaires.
11. https ://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

Python. [12]

Python est un logiciel de programmation né en 1991 par l'initiative de Guido van Rossum. C'est un logiciel qui est particulièrement utilisé dans les domaines de l'intelligence artificielle ou du traitement de la big data. [13] L'avantage de Python est qu'il possède énormément de bibliothèques qui lui permettent la réalisation de tâches variées. Pour le TAL, il y a deux bibliothèques qui sont généralement utilisées.

3.2.2.2 Présentation des bibliothèques Python

Natural Language Toolkit (NLTK) est la bibliothèque la plus populaire pour le traitement du langage naturel. Développée par Bird *et al.* (2009), elle possède beaucoup d'algorithmes qui lui permette la réalisation d'un grand nombre de tâches. Cette bibliothèque est souvent utilisée par les chercheurs car elle est polyvalente. D'un autre côté, Spacy, développé par Honnibal et Montani (2017) est une bibliothèque plus récente, capable comme NLTK de réaliser les principales tâches du TAL (tokenisation, lemmatisation, élimination des mots-vides...). Spacy possède de meilleures performances sur ces tâches que son concurrent, mais a aussi l'avantage d'être plus simple d'utilisation. Nous utiliserons donc Spacy pour rechercher les pratiques prescriptives.

3.2.2.3 Présentation de Spacy

Spacy est capable de faire ce genre de traitement automatisé par l'intermédiaire d'une pipeline. On peut définir ce terme par un ensemble de processus qui sont appliqués à des données, pour avoir en sortie des données traitées. Cela veut dire que l'ordinateur doit apprendre à identifier des données sans aide humaine. Pour cela, il faut récolter des données, les pré-traiter et entraîner la machine sur ces données. La machine est ensuite testée sur des données qu'elle ne connaît pas. Le choix des données va influencer sur les résultats que l'on peut avoir. Si les données sont variées, les résultats seront meilleurs, car ils couvriront un plus

12. https ://www.python.org/
13. Big data est un terme né à l'aube d'Internet et de l'explosion du volume des données. Il est utilisé pour parler des grosses quantités de données, de leur stockage ainsi que leur exploration, notamment sur Internet.

large échantillon de la langue. Spacy propose quatre niveaux d'entraînement de son pipeline :

— fr_core_news_sm
— fr_core_news_md
— fr_core_news_lg
— fr_dep_news_trf

Les trois premiers niveaux (fr_core_news_) correspondent respectivement aux petit (sm), moyen (md) et grand (lg) modèle d'entraînement de la pipeline. Le modèle sm sera plus rapide, mais aura moins de précision, alors qu'au contraire lg aura plus de précision, car il a plus de données d'entraînement, mais il sera plus long à appliquer. Le dernier modèle est basé sur les modèles transformer, évoqués précédemment lors de l'état de l'art. Ce modèle est plus efficace que les autres, mais est moins rapide. Il serait donc intéressant de regarder si la différence entre une petite et un grand pipeline est importante.

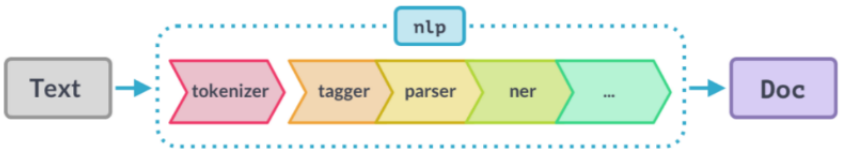

FIGURE 3.2 – Structure de la pipeline de Spacy

Pour illustrer le pipeline de Spacy, « Text » et « Doc » sont les données d'entrées et de sortie du pipeline. La section « nlp » répertorie les étapes du pipeline. Le schéma n'a pas présenté toutes les étapes, car il y en a trop pour les afficher. Dans les étapes présentes, on retrouve :

— *tokenizer* permet de découper le texte en tokens, qui correspondent ici aux mots.
— *tagger* correspond à l'étiquetage morphosyntaxique des tokens.
— *perser* identifie les relations de dépendance des tokens.
— *ner* va annoter les entitées nommées, c'est-à-dire l'ensemble des noms propres, comme des lieux, des individus ou encore des institutions.

Voici un exemple de la phrase « Je suis allé chez Apple. » après passage dans

le pipeline.

Tokenizer	**Tagger**	**Parser**	**NER**
Je	PRON	PRON	X
suis	AUX	AUX	X
allé	VERB	VERB	X
chez	ADP	ADP	X
Apple	PROPN	PROPN	ORG
.	PUNCT	PUNCT	X

TABLE 3.1 – Exemple de phrase traitée par la pipeline de Spacy

Pour la catégorie *NER*, « ORG » correspond à l'abréviation du mot « organisation ». Toutes les étiquettes morpho-syntaxiques proviennent du projet Universal Dependencies [14] (UD). Le projet construit des corpus arborés, autrement dit, ce sont des corpus annotés par les relations de dépendance que les mots ont dans une phrase. Le but de ses corpus est de faciliter l'entraînement des algorithmes d'analyse syntaxique. Le jeu d'étiquettes contient au total 17 catégories.

- ADJ: adjective
- ADP: adposition
- ADV: adverb
- AUX: auxiliary
- CCONJ: coordinating conjunction
- DET: determiner
- INTJ: interjection
- NOUN: noun
- NUM: numeral
- PART: particle
- PRON: pronoun
- PROPN: proper noun
- PUNCT: punctuation
- SCONJ: subordinating conjunction
- SYM: symbol
- VERB: verb
- X: other

FIGURE 3.3 – Jeu d'étiquettes morphosyntaxique de UD

14. https ://universaldependencies.org/u/pos/

Pour conclure cette partie, nous avons pu choisir notre outil pour effectuer notre mission. Nous avons observé en quoi Python et la bibliothèque Spacy étaient adaptés à notre travail. Grâce aux tâches dont dispose le pipeline de Spacy, nous pouvons effectuer des recherches contenant de plus riches informations linguistiques. La prochaine étape va présenter le prétraitement des données afin de faciliter la recherche de notre phénomène.

3.3 Prétraitement du corpus

Le prétraitement des données permet de passer de données brutes à des données transformées, afin de faciliter la recherche des pratiques prescriptives. Le traitement de données permet aussi d'alléger la charge de calcul de l'algorithme en enlevant les éléments que l'on considère comme inutile, ce qui est utile quand nous avons comme ici une grosse quantité de données à analyser. Les tweets contenus dans ce corpus n'ont pas subi de modification, mis à part l'anonymisation des pseudos et les adresses URL. Cependant, si on veut uniquement récolter des pratiques prescriptives, il faut enlever tous les éléments qui ne sont pas des mots ou qui ne rentrent pas dans le dictionnaire. Cela comprend les hashtags, les mentions de compte, les retweets, mais aussi les smileys.

Nous procédons à la suppression de ces éléments à l'aide de Regex, qui a une bibliothèque sur Python. Il permet d'effectuer des expressions régulières. Les expressions régulières sont un ensemble d'expressions qui, grâce à une syntaxe particulière, permettent de rechercher n'importe quelle chaîne de caractères. Voici l'expression régulière utilisée pour extraire les éléments cités précédemment :

1. *#\w+ ?|@XXX :? |RT |url_path| :|\w-|+ :*

L'expression régulière ci-dessus peut se découper en cinq parties, identifiée par les couleurs et séparée par « | » qui est l'opérateur logique OU. Elle permet de rechercher respectivement des hashtags, des mentions de compte, des retweets, des adresses URL ou des smileys. La suite de caractères bleue recherche un « # » suivi de chiffres, de lettres ou d'un « _ », répété au moins une fois (\w+) et qui peut se terminer par un espace (?), ce qui recherche des hashtags.

L'expression verte recherche les mentions de comptes. La fin de l'expression « :? » recherche la possible présence des deux-points, puis recherche un espace à la fin. L'expression rouge recherche simplement les caractères « RT » qui sont suivis d'un espace. « RT » est l'abréviation pour retweet et est présent quand une personne répond à un tweet en reprenant le tweet original. L'expression orange permet simplement d'identifier les adresses web, normalisées ici par « url_path ».

La difficulté ici a été de chercher les smileys, parce que ce ne sont pas des caractères classiques comme le sont les hashtags, les mentions, ou les retweets, mais des symboles plus complexes. Tous les emojis ont un standard Unicode [15] qui leur sont propre. Par exemple, le smiley ci-dessous à pour Unicode U+1F602.

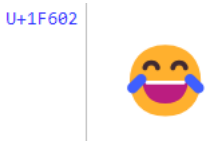

FIGURE 3.4 – Exemple d'un symbole avec l'Unicode associé

La recherche des smileys sur Python peut s'effectuer en recherchant tous les Unicodes des smileys. Cependant, cette tâche est assez complexe, car les smileys sont éparpillés dans la table Unicode. Une autre solution pour rechercher des émojis est de se servir de la bibliothèque Python « emoji ». Elle contient le consortium Unicode et permet de convertir les smileys en caractères identifiables et extractibles, grâce à la méthode *demojize*.

```
>>> print(emoji.demojize('Python is 👍'))
Python is :thumbs_up:
```

FIGURE 3.5 – Illustration de la méthode « demojize »

Le script ci-dessus permet de convertir le smiley de la chaîne de caractère représentée en rouge par des caractères classiques. Le format renvoyé est identique, on retrouve le nom du smiley en minuscule entouré par des deux-points.

15. Unicode correspond à un standard informatique dans lequel est stocké l'ensemble des caractères existants dans la majorité des langues.

C'est pour cela que la partie de l'expression régulière qui vise les smileys se compose de deux points entourés de lettres, de chiffres ou des caractères « _ - ».

Le mode d'écriture sur les réseaux sociaux est plus libre que sur les journaux par exemple, où aucune faute n'est acceptée. On peut supposer que la tolérance aux fautes n'est pas la même sur ces deux plateformes. Les utilisateurs obéissent sans doute à des règles tacites sur les réseaux sociaux, qui leur permettent de s'exprimer avec plus de liberté dans la communication verbale. Après ce travail de prétraitement des données, nous allons tester une méthode pour extraire notre phénomène.

3.4 Une extraction quantitative

3.4.1 Objectif de l'extraction quantitative

Notre problématique s'attarde sur la manière dont nous pouvons extraire les pratiques prescriptives. Notre objectif est de récolter le plus de phénomène possible. Si la méthode choisie n'est pas concluante, nous nous tournerons vers une méthodologie plus qualitative. Comme nous l'avons défini plus tôt dans le mémoire, un comportement prescriptif va juger une expression ou un terme linguistique en se référant à la norme d'une langue. C'est-à-dire qu'il est susceptible d'avoir des comportements prescriptifs dans ce qui est considéré comme à l'écart de la norme. D'après ce postulat, nous pouvons faire en sorte de chercher chaque élément qui est considéré comme une faute en français, puis de les extraire afin d'avoir une liste de fautes. Il faudra enfin trier ce qui est un cas d'usage ou bien un commentaire sur la faute extraite. C'est une méthode risquée, car premièrement, comme nous l'avons dit précédemment, le type d'écriture des réseaux sociaux ne respecte pas toujours la norme de la langue. La deuxième raison est que nous visons une grande quantité de données, ce qui implique d'être très précis dans l'extraction des fautes.

3.4.2 Description de la méthodologie quantitative

Même si notre objectif est d'avoir un maximum de pratiques prescriptives, il n'est pas possible de récolter toutes les occurrences, puisque ce terme englobe beaucoup de concepts différents. Comme nous allons identifier tous les éléments qui sont des fautes, nous nous intéressons ici au niveau du mot et donc aux fautes d'orthographe. La méthode choisie pour ce type d'extraction est la suivante : nous allons comparer chaque mot du corpus de tweets avec un lexique des mots français et nous y extrairons un fichier contenant tous les mots n'étant pas présent dans ce lexique.

3.4.2.1 Choix du lexique

La base de données « Lexique [16] » compte 140 000 mots de la langue française. Élaborée par New *et al.* (2004), cette base de données compte de nombreuses informations linguistiques sur les mots de la langue française. Ce lexique est distribué sous une licence non-commerciale (CC BY-NC 4.0), ce qui nous permet de l'utiliser librement sans utilisation commerciale. Autrement dit, il est interdit de produire de l'argent avec ce lexique. Le lexique comprend autant de mots, car il liste les flexions de tous les mots du français. Les flexions sont, en grammaire, l'ensemble des informations grammaticales qu'un mot peut avoir. C'est-à-dire que pour le mot « poucet » on compte 4 formes en associant le masculin, le féminin, le singulier et le pluriel (poucet, poucets, poucette et poucettes). Le lexique comporte d'autres informations comme la transcription phonétique en API [17] du mot, son lemme ou encore la classe grammaticale, mais nous utiliserons uniquement la colonne associée aux mots.

Le tableau 3.6 ci-dessous contient les colonnes citées précédemment : « Word » contient les mots, « phon » la trancsription phonétique, « lemme » la forme lemmatisée et « cgram » sa classe grammaticale.

16. http ://www.lexique.org/
17. L'API (alphabet phonétique international) est un alphabet utilisé pour transcrire les sons du langage parlé.

Word	phon	lemme	cgram
a	a	a	NOM
a	a	avoir	AUX
a	a	avoir	VER
a capella	akapEla	a capella	ADV
a cappella	akapEla	a cappella	ADV
a contrari	ak§tRaRjo	a contrari	ADV
a fortiori	afORsjoRi	a fortiori	ADV
a giorno	adZjORno	a giorno	ADV
a jeun	aZ1	à jeun	ADV

FIGURE 3.6 – Extrait de Lexique

3.4.2.2 Description du script

D'un point de vue technique, nous manipulons les tableaux à l'aide de la bibliothèque Python « Pandas »[18]. Cette bibliothèque permet de créer, parcourir et modifier des tableaux plus rapidement. Nous comparerons les tweets prétraités avec le lexique. Le script se base sur un extrait du corpus qui contient les 10 000 premiers tweets, pour faciliter la rapidité de l'exécution. La première étape du script consiste à extraire les deux colonnes qui contiennent les tweets ainsi que les mots du lexique.

Il n'est pas possible de comparer les colonnes avec leur état actuel, car Pandas compare les informations présentes dans chaque case, or les cases des tweets contiennent plusieurs mots qui les empêchent de pouvoir être comparés avec le lexique. Il est alors nécessaire de découper les tweets en mots. Python possède la méthode « .split() » permettant de séparer une chaîne de caractère en fonction du caractère voulu, l'espace est le caractère par défaut. En séparant les caractères par des espaces, nous nous retrouvons avec des mots, comme sur cet exemple :

2. *Il fait chaud aujourd'hui !*
['Il', 'fait', 'chaud', "aujourd'hui", '!']

Cependant, cette méthode de segmentation montre des limites, comme nous

18. https ://pandas.pydata.org/

allons l'observer. Ces mots sont ensuite intégrés dans une nouvelle colonne « mots », puis la colonne est comparée avec celle du lexique, pour garder les mots qui ne sont pas dans le lexique. La sortie de ce script est un fichier de type CSV, contenant une colonne avec les tweets et une autre avec le mot extrait du tweet qui n'est pas présent dans le lexique.

3.4.2.3 Limites de la récolte

Le fichier CSV contient bien des mots non-présents dans le lexique, mais possède énormément de données non souhaitées. Ce bruit [19] est beaucoup trop important pour récolter notre phénomène. Le script se base sur un extrait du corpus qui contient les 10 000 premiers tweets. Ce fichier a récolté plus de 27 000 mots, ce qui fait une moyenne de 2,7 fautes par tweets. Dans les données recueillies par le script, nous retrouvons différents types de mots ou caractères que nous allons classer dans le but de pouvoir les enlever du tableau. L'utilisation de la méthode .split() nécessite un nouveau traitement des données, car séparer les chaînes de caractère par des espaces ne garantit pas forcément de bien extraire un mot. Tout ce qui n'est pas un espace sera divisé puis comparé au lexique. Un mot suivi d'un signe de ponctuation sans espace sera extrait, même si ce mot est bien écrit.

— Ici, l'exemple 3 se différencie de l'exemple 2 car il n'a pas d'espace entre les ponctuations et les mots. Ce qui a pour conséquence de prendre en sortie les mots « chaud » et « aujourd'hui ».

3. *Il fait chaud, aujourd'hui !*
['chaud,', "aujourd'hui !"]

— L'ensemble des symboles qui ne sont pas utilisés dans le lexique du français sont à supprimer, car ils sont responsables d'une partie du bruit collecté. Dans ces symboles, on retrouve les nombres, les signes de ponctuation ainsi que quelques caractères spéciaux.

19. Le bruit correspond à l'ensemble des éléments récoltés par le script, mais qui ne sont pas souhaités.

4. *Ça represente meme pas 35% de son jeu, le mec montre zero but en 4min de vidéos, il a complié tout les rates de suarez...*
['35', '%','4min','suarez...']

5. *Ceux qui suive la NBA, Lakers ou Houston cette nuit ?*
[' ?','NBA,']

Dans les exemples 4 et 5, nous remarquons plusieurs termes qui ressortent. Les chiffres, qu'ils soient seuls ou suivis d'un mot, mais aussi les caractères tels que les pourcentages, le point d'interrogation ou la virgule sont extraits. Par ailleurs, le terme « NBA » aurait tout de même été séléctionné sans la virgule qui lui succède, nous allons voir pourquoi.

— La normalisation apportée au texte porte un autre problème, car les majuscules sont toujours présentes, ce qui fait que le lexique ne va pas reconnaître un mot bien orthographié s' il contient une majuscule. De plus, les mots contenant des majuscules peuvent être des entités nommées, mais sont présents dans le tableau, bien que ce ne soient pas des fautes.

6. *Ceux qui suive la NBA, Lakers ou Houston cette nuit ?*
['Ceux','NBA,',' Lakers,','Houston']

Les exemples 5 et 6 sont identiques, mais les mots récoltés ne sont pas les mêmes pour mettre l'accent sur les différentes catégories. Ici, l'exemple 6 prend les trois entités nommées, mais aussi tous les mots contenant des majuscules, comme les mots en début de phrase.

— L'élision est un phénomène linguistique qui va supprimer une lettre devant une consonne afin de faciliter la prononciation du mot. Ce n'est pas une faute, mais le lexique ne contient pas la forme élidée de chaque mot. Il est donc nécessaire d'enlever ces formes, au risque que l'algorithme extrait des mots grammaticalement corrects, comme dans l'exemple 7.

7. *Maroua a chaque drama elle me déçoit un peu plus je suis choquée*

d'elle
['d'elle']

— Nous retrouvons également des abréviations dans les tweets. Certaines ne sont pas considérées comme des fautes. L'exemple 8 est une entité nommée, mais aussi une abréviation de « National Basketball Association ». Les abréviations ne sont pas toujours des entités nommées comme l'illustre l'exemple 9.

8. *Ceux qui suive la NBA, Lakers ou Houston cette nuit ?*
'NBA,'

9. *mdr vraiment bizarre de spoil l'arrivée de Sabo comme ca (si c vraiment le cas)*
['mdr']

— Enfin, nous trouvons parfois des mots d'une autre langue ou bien des anglicismes, c'est-à-dire des emprunts à la langue anglaise avec des modifications linguistiques, qui peuvent être de l'ordre sémantique ou morphologique entre autres.

10. *mdr vraiment bizarre de spoil l'arrivée de Sabo comme ca (si c vraiment le cas)*
['spoil']

3.4.2.4 Modification du pré-traitement des tweets

Cette typologie nous sert à comprendre pourquoi l'algorithme n'a pas récolté les données voulues, puis nous permet de cerner leur caractéristiques pour modifier notre script et supprimer ces éléments inutiles. Nous devons ajouter des informations à l'expression régulière qui permet de pré-traiter les tweets.

11. `\w+['']`
`[^A-Za-zïêëâüçûôéèñàãöùî ]+`
`d+\S*`

L'expression en première position va rechercher les mots élidés (l', d', puisqu'...). Celle en deuxième recherche l'ensemble des symboles qui ne sont pas utiles à l'analyse de l'orthographe d'un mot, comme nous l'avons vu dans l'exemple 3 et 4. Nous avons recherché tous les caractères présents dans le lexique. Afin d'être plus précis dans la recherche, nous avons parcouru l'ensemble des mots du lexique et avons gardé les caractères uniques, qui correspondent à l'expression. Enfin, la dernière expression recherche le ou les chiffres, ainsi que ce qui suit les chiffres. Par exemple, « 45min » doit être récolté dans son entièreté, car « min » est une abréviation, or le lexique n'en contient pas.

3.4.2.5 Limites de la méthode et axes d'amélioration

Malgré les modifications apportées à l'expression régulière, on retrouve encore énormément de données en sortie (25 331 lignes). Les expressions régulières montrent leurs limites et ne permettent pas de supprimer toutes les données qui nous dérangent. Dans l'exemple 6, il serait possible d'enlever tous les éléments qui possèdent des majuscules, mais tous les mots en majuscule ne sont pas forcément des entités nommées. Nous retrouvons par exemple des mots écrits en majuscule, ou encore les majuscules en début de phrase. Nous avons fait le choix de mettre les tweets en minuscule pour enlever les mots correctement orthographiés qui sont en majuscule. Voici un extrait du tableau de sortie :

tweet_id	tweet	mot
82013_tweet.txt	bosh il lui a fait un sal boulot à kaaris il lui a montré que maintenant a plus de petit et grand	bosh
82013_tweet.txt	bosh il lui a fait un sal boulot à kaaris il lui a montré que maintenant a plus de petit et grand	sal
82013_tweet.txt	bosh il lui a fait un sal boulot à kaaris il lui a montré que maintenant a plus de petit et grand	kaaris
130701_tweet.txt	le craquage est complet ai préco le repackage de straykids et mario allstar	préco
130701_tweet.txt	le craquage est complet ai préco le repackage de straykids et mario allstar	repackage
130701_tweet.txt	le craquage est complet ai préco le repackage de straykids et mario allstar	straykids
130701_tweet.txt	le craquage est complet ai préco le repackage de straykids et mario allstar	mario
130701_tweet.txt	le craquage est complet ai préco le repackage de straykids et mario allstar	allstar
98773_tweet.txt	chut faut pas le dire à camille	camille
258184_tweet.txt	ma passion pendant le générique des vlogs août essayer de deviner qui est qui quand a les dessir	vlogs
258184_tweet.txt	ma passion pendant le générique des vlogs août essayer de deviner qui est qui quand a les dessir	léna
262936_tweet.txt	me encontré bts	encontré
262936 tweet.txt	me encontré bts	bts

FIGURE 3.7 – Tableau de récolte des fautes du corpus

Comment expliquer ces résultats si peu concluants ? L'erreur ici est d'abord de choisir une méthode s'axant sur une vérification orthographique alors que les données ne s'y prêtent pas, puisque d'abord, le registre de langue est familier

dans les tweets choisis. Le registre familier possède plus de libertés linguistiques. Il serait intéressant de regarder si les tweets avec un registre plus soutenu ont une proportion de mots récoltés plus basse. Nous pourrions aussi supposer que les pratiques prescriptives sont présentes chez des personnes qui ont un registre de langue plus soutenu. Analyser l'orthographe des tweets s'avère moins cohérent, comme nous l'avons experimenté. Cette méthode serait plus adaptée à des données journalistiques par exemple.

Ensuite, le postulat de départ qui était que toutes les fautes sont de valeur prescriptive n'est pas correct. Il est plus correct de dire que certains termes ont un potentiel prescriptif plus important. C'est-à-dire qu'ils ont plus tendance à être soumis à des pratiques prescriptives. Les raisons peuvent varier d'une expression à l'autre, mais généralement, plus un terme n'est pas employé dans la norme, plus on a de chance de trouver des commentaires prescriptifs.

Il serait toujours possible de trouver une méthode pour extraire les entités nommées et les abréviations ; Spacy possède bien un outil pour extraire les entités nommées, mais ces dernières sont tellement variées dans le corpus qu'il n'a pas été entraîné à tous les récolter. Nous pourrions aussi prendre un corpus qui liste des entités nommées et des abréviations, mais le problème demeure identique.

Pour conclure, la méthode quantitative nous montre la difficulté de rechercher un phénomène aussi vaste qu'une pratique prescriptive. Nous pouvons noter aussi qu'il n'est pas forcément nécessaire de vouloir récolter le plus de données possible si ces dernières ne sont pas pertinentes pour répondre à la problématique. Il faut donc être plus précis pour avoir des résultats concluants, quitte à ne pas pouvoir tout récolter. La seconde méthode va s'axer sur une recherche plus qualitative de ces pratiques prescripives.

3.5 Une extraction qualitative

3.5.1 Méthodologie qualitative des expressions à tendance prescriptive

Après avoir testé une méthode quantitative pour récolter les pratiques prescriptives, nous allons regarder comment procéder pour récupérer des pratiques prescriptives de manière plus précise. Cette méthode est plus simpliste, car nous allons uniquement sélectionner certains phénomènes. L'idée est de limiter le bruit récolté en gardant seulement des phénomènes qui ont un plus grand potentiel prescriptif.

3.5.1.1 Choix des termes prescriptifs

Le choix de ces termes peut se faire en fonction de la fréquence. Plus un phénomène revient souvent, plus il a de chance d'être soumis à une critique prescriptive. Le site de l'Académie française liste les bons et mauvais emplois de la langue française dans la section « Dire, Ne pas dire » [20]. Dans ces emplois, nous allons choisir de prendre l'expression « comme même » ainsi que les auxiliaires suivis d'un verbe à l'infinitif et vice versa. Par exemple, on pourrait récolter les phrases « Il a manger.* » et « Il peut mangé.* ». Les fautes de conjugaison, notamment l'inversion du passé composé et des verbes du premier groupe de l'infinitif sont assez populaires. On suppose que l'homophonie de ces deux formes y joue pour beaucoup. L'expression « comme même » à la place de « quand même » est assez présente dans le langage courant. Nous recherchons les termes sans prétraiter le corpus, pour éviter d'enlever des informations qui seraient utiles.

3.5.1.2 Methodologie de l'extraction des expressions

L'extraction de l'expression « comme même » est plus simpliste puisque nous effectuerons un requêtage dans le corpus pour directement trouver les occurrences de l'expression. Nous exécuterons cette méthode sur tout le corpus,

20. https ://www.academie-francaise.fr/la-langue-francaise-dire-ne-pas-dire/dire-ne-pas-dire

contrairement à la méthode quantitative. Le but est de trouver le plus d'occurrences possibles. La seule différence que nous faisons est la présence, ou non, de l'accent circonflexe. Nous considérons que la valeur prescriptive se portera plus sur l'expression en soi que sur cette faute d'orthographe. L'expression se présente comme suit :

12. *comme m[êe]me*

Après recherche du terme dans le corpus, le très faible nombre d'occurrences nous oblige à regarder d'autres phénomènes en détails afin d'étoffer notre récolte. La fonction « Rechercher et Remplacer » du logiciel Excel nous suffit pour regarder l'ensemble des occurrences d'un ensemble de caractères. Après avoir testé un ensemble d'expressions à valeur prescriptive, voici les mots qui ont été séléctionnés ainsi que leur expression régulière pour les extraire via Python :

Expression	**Expression régulière**
comme même	comme m[êe]me
iel	[\n]iel[\n]
croive/ent	croive
voye/ent	voye[nt]+
malgré que	malgré qu

TABLE 3.2 – Séléction des expressions à valeur prescriptive

Le mot « iel » est assez controversé et est donc soumis à des critiques prescriptives. Son expression régulière recherche le mot entouré d'un espace ou d'un retour à la ligne. En indiquant seulement les caractères du mot, on peut récolter du bruit qui est en grande partie des mots contenant iel (le mot « événementiel » par exemple). Les expressions « croivent » et « voient » ont déjà été étudiées chez Humphries (2023). L'expression régulière « voyent » permet de ne pas prendre la forme de voir au présent de l'indicatif, à la 2e personne du pluriel (voyez). Enfin « malgré que » est une expression assez courante, son expression régulière ne contient pas de « e » pour rechercher aussi sa forme élidée.

L'extraction des verbes se fera grâce à Spacy. Ici, il n'est pas possible de rechercher ces phénomènes uniquement avec des expressions régulières. Elles ne permettent pas de rechercher une classe grammaticale comme nous voulons le faire. Spacy permet d'identifier les classes grammaticales des mots. Grâce à cela, nous allons identifier les endroits où un mot est identifié comme auxiliaire et un suivi d'un verbe à l'infinitif pour le premier cas de figure. Pour le second, nous allons regarder les verbes conjugués au participe passé, mais n'ayant pas d'auxiliaire qui le précède. Le risque ici est de manquer des phénomènes, car l'organisation syntaxique d'un verbe n'est pas toujours dans l'ordre « auxilliaire - verbe ». Nous pouvons par exemple trouver un adverbe entre ces deux composants.

13. *Il* ***a échanger*** *son savoir.**
Il ***a*** *souvent* ***échanger*** *son savoir.**
*Il n'****a*** *pas souvent* ***échanger*** *son savoir.**

Les verbes de l'exemple 13 sont indiqués en gras. On note qu'avec la méthode choisie, en l'état, il serait seulement possible de détécter la première phrase. La seconde phrase contient un adverbe et la troisième en contient même deux. Il faut alors aussi prendre en compte ces cas.

Pour commencer, nous allons rechercher les erreurs de conjugaison au passé composé (« -er » au lieu de « -é »). Pour cette méthode, nous allons utiliser l'annotation morphosyntaxique de Spacy afin de rechercher chaque mot et garder ceux étant des verbes. Nous allons ensuite rechercher la terminaison du verbe en utilisant la méthode .endwith, puis d'indiquer les verbes se terminant par « -er ». A cette étape, il faut encore identifier une potentielle présence d'auxiliaire dans la phrase. La solution choisie est de regarder dans le contexte précédent le verbe. Si le lemme « être » ou « avoir » n'est pas présent, c'est-à-dire un auxiliaire, alors nous extrayons le verbe. La méthode concernant les erreurs de conjugaison à l'infinitif (« -é » au lieu de « -er ») est identique, mais au lieu de prendre le verbe quand il n'y a pas d'auxiliaire, on le prend quand un auxiliaire est détécté dans le contexte précédant le verbe.

Le fichier de sortie de toutes ces expressions est un fichier CSV contenant

quatre colonnes. La première colonne « type » indique si l'expression vient d'une des expressions régulières du tableau 3.5.1.2, est un verbe avec une terminaison en « -é » ou un verbe avec une terminaison en « -er ». La seconde colonne « id » reprend les identifiants liés aux tweets, la troisième colonne « expression » indique l'expression extraite et la dernière colonne « tweet » affiche le tweet provenant de l'expression. Voici un extrait du tableau de sortie :

type	id	expression	tweet
-é	262936_tweet.txt	encontré	Me encontré #MTVHottest BTS @XXX url_path
-er	160768_tweet.txt	frapper	#Hassan les personnes qui croient encore que Hassan a frapper fathma venez on se pose 2 min e
-é	58099_tweet.txt	manqué	Sa soule trop de changement sur #rmclive #RMCF après bourdin,brunet maintenant fatima j aime
-é	5161_tweet.txt	publié	@XXX Moi c'était zaza mimosa j'étais trop deg de savoir que cetait plus publié ??
-é	241445_tweet.txt	affamé	Plus qu'un tour, Sainz affamé !!! #ItalianGP #F1 #AuRupteur url_path
-é	240193_tweet.txt	partagé	Par conte partagé pas des photos ou vidéos entrain de decapiter l'animal svp j'ui sensible
-er	56768_tweet.txt	commencer	@XXX Et ces Bucks HORRIBLE a commencer par leur leader et leur Coach
-er	47334_tweet.txt	écouter	Bon bah, on va commencer a écouter du freeze, le colors show était lourd
-é	115806_tweet.txt	préféré	@XXX Ah vraiment c'est dommage j'aurais préféré OKC , les Lakers ils vont s'amuser
-er	245276_tweet.txt	trouver	Semaine prochaine si j'ai pas trouver de maillot de bain j'me tuz
-é	3422_tweet.txt	café;Liberté	Vendredi 4 septembre ; twi twi twi venez mes amis twittos prendre un tweet-café;Liberté Egalité,
-é	3422_tweet.txt	Fraternité	Vendredi 4 septembre ; twi twi twi venez mes amis twittos prendre un tweet-café;Liberté Egalité,
-é	140887_tweet.txt	créé	Vous êtes vraiment culottés ici. Vous osez dire Ouais qu'ils s'occupent de faire monter leur fav au
-é	140887_tweet.txt	créé	Vous êtes vraiment culottés ici. Vous osez dire Ouais qu'ils s'occupent de faire monter leur fav au
-er	208006_tweet.txt	danger	@XXX Avec Lacazette en 10 ou ozil non franchement c'est danger

FIGURE 3.8 – Tableau des expressions à valeur prescriptive

Cependant, les données en sortie ne sont pas toujours satisfaisantes. Au total, le ficher contient 6 440 expressions. Nous verrons lors du prochain chapitre le détail des données colléctées.

Pour conclure cette partie, nous avons pu collecter avec plus de précision des fautes, qui nous permettent de chercher par la suite les pratiques prescriptives. Nous avons cependant remarqué qu'il y a encore beaucoup de données collectées, ce qui rend compliquée l'analyse manuelle de toutes ces expressions. Nous allons tester une dernière méthode, similiaire à celle-ci dans sa méthodologie, mais qui ne recherche pas de fautes.

3.5.2 Méthodologie qualitative des expressions prescriptives

3.5.2.1 Présentation de la méthode

Depuis le début de ce mémoire, nous recherchons des fautes dans le but de trouver des pratiques prescriptives. Cependant, la plupart des fautes sont produites par des locuteurs sans pour autant que la faute soit critiquée explicitement. Il faudrait pour cela accéder aux réponses des tweets afin d'espérer trouver une réponse qui serait prescriptive, mais cette méthode s'avérerait chronophage. Une solution pour trouver des pratiques prescriptives est simplement de rechercher des expressions prescriptives. La différence avec la méthodologie précédente est qu'ici, nous cherchons directement la pratique prescriptive et non un terme qui pourrait amener à une pratique prescriptive.

14. *Mais qu'est ce que tu raconte tu préfères qu'on te definise a ta couleur de peau plutôt qu'a ta personnalité c'c'est idiot le terme* ***racisé*** *est immonde selon moi parce qu'elle fait une différence entre les individus*

Pour l'exemple 14, l'expression prescriptive est ici « le terme racisé est immonde », avec « racisé » comme terme visé par la critique. Comme pour la recherche des expressions à tendance prescriptive, il faut savoir de quelle manière nous allons rechercher ses expressions. Pour cela, nous pouvons nous aider des expressions prescriptives annoteés lors de mon stage. Cela nous donnera une idée des expressions qui peuvent être utilisées.

3.5.2.2 Description de la méthode

La méthode choisie pour extraire ces termes est la même qui est utilisée pour extraire les expressions régulières du tableau. 3.5.1.2. Après avoir testé différentes expressions, voici celles qui sont choisies pour être extraites :

Expression	**Expression régulière**
orthographe	orthographe
langage	langage
on dit	\bon dit(?!\squ)
le terme/mot	le (terme|mot)

TABLE 3.3 – Séléction des expressions prescriptives

Nous choisirons ces quatre expressions afin d'extraire les pratiques prescriptives. Elles ont été choisies parmis un ensemble de termes qui ont été requêtés via Excel et sa fonction rechercher et remplacer. Le but était de trouver des expressions qui avaient tendance à attirer des conversations métalinguistiques. « orthographe » et « langage » sont directement liés au langage et donc sont assez présents dans les expressions préscriptives. L'expression régulière des deux termes est identique, on recherche simplement les caractères des mots. L'expression « on dit » est de nombreuses fois dans des contextes prescriptifs. L'expression régulière recherche l'expression qui n'est pas suivie des caractères « qu », car les occurrences « on dit qu-e » sont très présentes et ne sont pas des pratiques prescriptives. Enfin, les expressions « le terme/mot » recherchent ses deux synonymes. Ils sont généralement présents avant de parler d'un terme. Son expression recherche l'article suivi du mot « terme » ou du terme « mot ». L'article vient mettre l'accent sur le mot qui est normalement cité ensuite.

Nous avons écarté certains termes comme « langue », qui se rapproche de « langage » mais qui souffre de sa polysémie, puisque beaucoup d'occurrences proviennent d'un contexte qui sont des expressions figées ou de l'organe, comme l'illustrent ces deux exemples :

15. *RT @XXX : Trop hâte de me marier et de faire taire toute c petite langue de vipère*
Vous aussi vous avez des tics ? Si vous lesquels ? Perso quand je suis grave concentré sur qlq je mords ma langue mdr

Le mot « français » ne contient pas d'occurrences prescriptives, car là aussi, les sujets ne se tournent pas vers le métadiscours, mais plutôt vers la politique

comme le montre l'exemple 16. Les résultats ne sont pas plus concluants en combinant les termes.

16. *#Immigration est à bannir ! Dehors ceux qui ne sont pas francais ou européens !*

Le fichier de sortie est un fichier CSV avec trois colonnes. On retrouve les mêmes colonnes que la deuxième méthode, c'est-à-dire « id » qui contient la clé d'identification du tweet, « expression » qui affiche le terme que l'expression régulière à extrait et « tweet » qui contient le contenu textuel du tweet. Le fichier de sortie contient 600 lignes.

id	expression	tweet
7943_tweet.txt	on dit	@XXX @XXX @XXX @XXX à moins de soliloquer , on dit on étouffe #ggrmc
24187_tweet.txt	on dit	@XXX @XXX @XXX @XXX @XXX @XXX @XXX @XXX @XXX Ca fait quand même beaucoup. C
91808_tweet.txt	On dit	On dit toujours Mister Freeze mais les Yeti c'est large au dessus url_path url_path
143790_tweet.txt	On dit	On dit "calabraise testa dura" , ou on dit "j aurais dû fermer ma gueule" #rmclive
137790_tweet.txt	on dit	@XXX Ah voilà j'ai répondu pareille on m'a pris pour une fadade chez nous on dit Yeti ??????
194350_tweet.txt	on dit	@XXX @XXX @XXX Oh pauvre bête biquette, pas contente parce qu'on dit la vérité sur la racaille a
108087_tweet.txt	on dit	@XXX La juge est décédé ou quoi c'est n'importe quoi il a pas fais exprès quand djoko casse c raq
207460_tweet.txt	on dit	La chose chelou avec histoire de Ademo c'est que bcp de média ou personne de la télé on dit bie
218310_tweet.txt	on dit	@XXX @XXX @XXX @XXX @XXX @XXX @XXX Bon allez à 3 on dit tous une dernière fois #AdamaVi
220536_tweet.txt	on dit	@XXX @XXX @XXX @XXX LOOOOOOL, c'est connu les suceur de tootatis c'est tout s que vous ave
74487_tweet.txt	on dit	@XXX jveut trop visiter roubaix pr voir c'si c guez comme on dit
29881_tweet.txt	ON DIT	@XXX JE PARLE DES GENS QUI ONT TAILLER ADELE LES GENS QUI ON DIT DE BOYCOTTE DOJA MA
63799_tweet.txt	on dit	C'est comme quand on dit : Liebig con comme ses pieds ! C'est une image ... ?? #ggrmc @XXX
229001_tweet.txt	on dit	@XXX @XXX Mdr il on dit d'eux fous tu sais lire ? Danilo c'est une fois

FIGURE 3.9 – Tableau des expressions prescriptives

Pour conclure, la méthode de recherche directe des expressions prescriptives est sans doute celle qui répond le mieux à notre problématique. En effet, elle nous permet au maximum un tri manuel des données et maximise le taux de phénomènes dans le fichier de sortie, puisque nous avons ciblé les phénomènes et non des termes qui pouvaient abriter des occurrences avec des pratiques prescriptives. Cependant, cette méthode se limite à certaines expressions, cela implique que l'on passe à côté de phénomènes présents dans d'autres contextes.

3.5.3 Conclusion du chapitre

Lors de ce chapitre, nous avons décrit l'ensemble du processus réalisé pour répondre à la problématique. Notre corpus a été choisi dans l'objectif d'avoir un maximum d'internautes non-experts de la langue, ce que les réseaux sociaux contiennent en abondance. Nous avons ensuite choisi Python comme outil de recherche et d'extraction des pratiques prescriptives, pour sa polyvalence et sa grande collection de bibliothèques, notamment Spacy pour le TAL. Après avoir réalisé un prétraitement sur le corpus, nous avons testé trois méthodes permettant d'extraire des pratiques prescriptives du corpus. Les méthodes sont téstées du plus au moins quantitatif. La première méthode avait pour ambition de récolter le plus de phénomènes possibles en identifiant les fautes présentes dans les tweets. Les résultats n'étant pas convaincants, nous avons opté pour une méthode plus précise qui consiste à cibler des termes qui ont tendance à attirer les phénomènes. La dernière méthode vise directement les expressions prescriptives, en cherchant des expressions choisies au préalable dans le corpus.

La première méthode semble être celle qui est la moins efficace, car elle récolte beaucoup trop de bruit, en plus d'extraire des données qui ne sont pas pertinentes. La seconde méthode semble être plus précise, mais la dernière méthode paraît être la meilleure, car elle n'identifie pas les fautes comme les deux autres méthodes, mais directement les expressions. Le prochain chapitre s'attarde à analyser les résultats obtenus des trois méthodes, ce qui nous permettra d'apporter une réponse à nos hypothèses, puis de répondre à notre problématique.

Chapitre 4

Analyses et résultats des méthodes de récolte

L'objectif de ce chapitre est de comprendre les résultats obtenus en analysant les donneés des trois méthodes. À l'issue de ces analyses, nous pourrons choisir la méthode qui est la plus performante dans l'extraction de pratiques prescriptives.

4.1 Mesures de précision et de rappel

Les mesures de précision et de rappel vont nous permettre d'analyser ces résultats. La précision calcule la part des items valides sur l'ensemble des items. Le rappel calcule la part des items valides sur l'ensemble des items valides. Voici les deux mesures ci-dessous :

$$Précision = \frac{Vrai\ positifs}{Vrai\ positifs + Faux\ positifs} \qquad \frac{15}{15+25} = 40\%$$

$$Rappel = \frac{Vrai\ positifs}{Vrai\ positifs + Faux\ négatifs} \qquad \frac{15}{15+10} = 60\%$$

FIGURE 4.1 – Mesure du rappel et de la précision

La F-mesure est la moyenne de ces deux mesures. Elle se calcule grâce à une matrice de confusion. Sa formule est 2*(rappel*précision)/(rappel+précision).

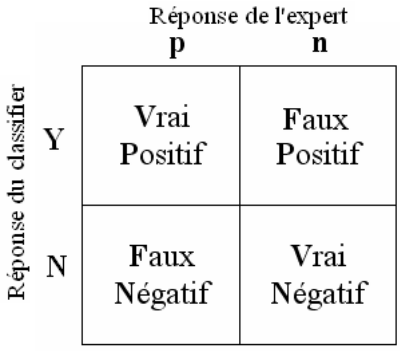

FIGURE 4.2 – Matrice de confusion

La part des items pertinents extraits s'appelle les vrais positifs, c'est le cas quand les items collectés correspondent à nos attentes. Au contraire, la part des items non-pertinents extraits sont les faux positifs. Les faux négatifs sont l'ensemble des items non pris en compte, mais qui sont pertinents. Au contraire, les vrais négatifs sont les items non-pertinents qui n'ont pas été détectés.

Pour illustrer cela, si on recherche des spams dans une boite mail, on cherchera à séparer les spams des non-spams. Par exemple, la sortie nous renvoie 100 mails, dont 80 sont des spams. Il y a donc 80 vrais positifs et 20 faux positifs, la precision est alors de 80/(80+20)=0,8. Dans les mails, l'algorithme a manqué 15 mails, ce qui fait 15 faux négatifs. Le rappel est donc de 80/(80+15)=0,84. On peut mesurer le f-mesure, qui donne : 2*(0,8*0,84)/(0,8+0,84)=0,82.

4.2 Analyse de la méthode quantitative

4.2.1 Analyses quantitatives de la première méthode

La quantité de mots récoltés avec cette méthode est de 25 331. Ce qui est toujours trop pour l'échantillon de 10 000 tweets sur lequel est testé le script. Cela représente environ 2,5 fautes par tweets. Pour être plus précis, nous pouvons calculer le taux de fautes par mots présents dans l'ensemble des tweets. La méthode « .shape » de Pandas nous permet de trouver un total de 178 784 mots. Cela représente une proportion de 14% de fautes. Sans l'étape du prétraitement des données, nous récoltons 72 786 lignes, contre 25 331 avec, soit près de 3 fois

plus de données. Cela représente même l'équivalent de 40% de l'ensemble des mots de l'échantillon du corpus.

Il n'est pas possible d'analyser l'ensemble des données. C'est pour cela que nous allons analyser les 200 premières lignes du tableau de sortie. Nous annoterons la présence ou non d'une pratique prescriptive, mais aussi le type de mots qui a été récolté. Dans les types de mots, nous retrouvons les abréviations, les anglicismes, les interjections, les néologismes, les noms propres, les fautes et les mots correctement orthographiés, mais présents dans la liste, voici le tableau qui présente ces catégories :

Etiquette	**Definition**	**N°**	**Exemple**
ABR	Les mots sont des abréviations.	44	On a notre match face à sochaux **tkt lol**
ANG	Les mots présents sont des anglcismes, ou des mots anglais.	25	Dommage le **feat** bosh Kaaris il est court
FAUTE	Les mots sont des fautes.	29	les stmg sont en tt et bah je ris toute seule sur tous les tweets depuis **talheur**
INTER	Les termes sont des interjections.	2	Ça promet contre l'Inter et le Bayern @XXX allez **ooooh**!!!
NEO	Les mots sont des néologismes ou des mots d'argot.	10	@XXX Je pense pas que je paniquais j'aurais juste bien le **seum**
NP	Catégorise les noms propres	83	**Leris Luketo** il a lâché une bombe et il est parti comme ça ??
X	Ces mots sont bien orthographiés, mais sont quand même présents.	7	les **journalistes/Consultants** en France c'est vraiment quelque chose !

TABLE 4.1 – Typologie des mots extraits

Pour cette méthode, les items pertinents sont l'ensemble des éléments qui ne

sont pas dans le lexique. C'est le cas pour tous les éléments, à l'exception de la catégorie X. La précision pour la première méthode est donc de 193/200=0,965. Le rappel prend ces 193 éléments pertinents par rapport à l'ensemble des éléments qui sont pertinents. Nous vérifions cela manuellement sur les 200 tweets. Nous n'avons pas trouvé d'autres mots pertients. Cela signifie que le rappel est parfait (193/193=1) Le f-mesure est donc de 2*(1*0,965)/(1+0,965)=0,98 ce qui est un résultat très performant.

Cependant, les items pertinents ici peuvent être les fautes, puisque notre objectif est de trouver des fautes. Dans ce cas, les noms propres (NP) et les faux positifs (X) ne sont pas des fautes, le reste des annotations sont en dehors de la norme et peuvent donc être considérés comme des fautes. Nous avons alors 110 vrais positifs. La précision donne 110/200= 0,55. Le rappel, c'est-à-dire la totalité des fautes présentes dans l'échantillon, est aussi parfait (110/110=1). Le f-mesure nous donne 2*(1*0,55)/(1+0,55)=0,71. C'est un résultat médiocre, car même si tous les vrais positifs ont été extraits, le surplus de bruit fait baisser le résultat.

Si l'on considère que seul la catégorie « fautes » possède une valeur prescriptive et est donc pertinente, alors la précision baisse même à 0,15 (29/200). Le rappel est logiquement aussi de 1. Le f-mesure est de 2*(1*0,15)/(1+0,15)=0,26. C'est un résultat insuffisant, qui s'explique aussi par une grosse quantité de bruit. Le tableau 4.2 récapitule les trois mesures. La catégorie « type » classe les trois types de valeurs pertinentes dont nous avons discuté. Le type « Fautes+ » comprend l'ensemble des catégories, à l'exception de NP et X.

type	**Précision**	**Rappel**	**F-mesure**
Lexique	0,965	1	0,98
Fautes+	0,55	1	0,71
Fautes	0,15	1	0,26

TABLE 4.2 – Rappel, Précision et F-mesure de la méthode n°1

4.2.2 Analyses qualitatives de la première méthode

L'échantillon de mots ne contient aucun tweet à valeur prescriptive. Sur les 200 mots analysés, 81 sont des noms propres, ce qui fait 41,5% des mots. C'est la catégorie qui est la plus présente, suivi par les abréviations avec 44 occurrences (22%). La catégorie « FAUTE » comprend 29 occurrences, soit 14,5% du total. Les anglicismes suivent de près avec 12,5% des mots (25 occurrences). On retrouve 10 mots de la catégorie « NEO » dans cet échantillon (5%). Les mots correctement orthographiés et quand même présents dans le tableau sont minoritaires (« X »), représentant 3,5% du total. Les deux derniers mots sont des interjections.

Dans les 29 occurrences de la catégorie faute, on peut retrouver des oublis ou des ajouts d'accents (« poele »), des élisions (« jferais »), des fautes d'accord (« ils demandent du travails ») et des fautes d'orthographe (« Bosh il lui a fait un sal boulot à kaaris [...] »). Pourquoi le script a récolté des éléments qui sont dans le dictionnaire ? Si on regarde plus en détail, nous remarquons de suite que l'erreur se trouve au niveau du prétraitement des données. Pour l'exemple « + Perso [...] simplement parce-qu'on vous demande d'en porter 1 en ville dans les magasins [...] », le prétraitement a supprimé les caractères « -qu' » et fusionné « parce » et « on » en « parceon ». Ce phénomène apparaît quand les caractères spéciaux sont entourés par deux mots sans espace. Il serait possible d'éviter cela en perfectionnant l'expression régulière.

4.2.3 Conclusion de la première méthode

Comme nous l'avons supposé, la méthode quantitative n'est pas pertinente et efficace pour trouver des pratiques prescriptives. L'échantillon montre même qu'aucune pratique prescriptive n'est présente sur 200 mots. Nous avons observé une grosse présence d'entités nommées qui augmente considérablement le bruit. Les fautes sont des cas d'usage du mot, elles ne prescrivent pas le mauvais usage du mot. Un échantillon plus large permettrait d'avoir des pratiques, mais les résultats n'en resteraient pas moins faibles.

4.3 Analyse de l'extraction des expressions à tendance prescriptive

4.3.1 Analyses quantitatives de la seconde méthode

Pour commencer, le fichier de sortie a extrait 20 292 expressions sur l'ensemble du corpus. Les résultats semblent similaires à la méthode précédente, mais cette dernière a été entraîné sur l'échantillon de 10 000 tweets. Si on compare à la même échelle grâce à un produit en croix, nous trouvons 888 tweets, soit près de 23 fois moins de données extraites en comparaison aux premières analyses.

type	**total**	**échantillon**	**%**
ex	62	62	100%
-er	6 616	41	0,6%
-é	13 616	97	0,7%
TOTAL	20 292	200	/

TABLE 4.3 – Répartition des types d'expression

Comme pour l'analyse précédente, nous avons séléctionné un échantillon de 200 tweets, soit environ 1% du corpus (20 292/200). Étant donné la faible portion du type « ex », nous avons décidé de tous les analyser. Quant aux deux autres, nous avons réparti les 138 tweets en fonction de leur part dans le tableau. 70% des tweets proviennent des termes ayant un suffixe « -é », ce qui donne 97 termes, le reste des 41 termes est attribué aux termes avec le suffixe « -er ». La colonne « % » affiche la part des termes présents dans le corpus. On remarque que les deux derniers types ne sont pas très représentés.

Nous allons maintenant calculer les scores de rappel et de précision pour ces données. Sur les 200 tweets, 91 sont considérés comme non-pertinents. La catégorie « ex » ne contient aucune expression non-pertinente. Ce qui donne un score de précision parfait (62/62=1). La catégorie « -é » est celle qui en possède le plus avec 74 expressions non-pertinentes. En conséquence, cela donne une précision de 0,24, qui est un mauvais score. La catégorie « -er » possède une

précision plus haute que la dernière catégorie, mais quand même assez moyenne : 0,59. Au total, cela donne une précision moyenne de 0,55.

La mesure de rappel est calculable en prenant en compte l'ensemble des éléments pertients, qui peut être identifiable manuellement en regardant si l'algorithme a bien extrait tous les éléments pertinents (Par exemple, on regarde si les expressions à l'infinitif précédé d'un auxiliaire sont toutes extraites). À partir de cette définition, « ex » possède encore une fois un score parfait (62/62=1). « -é » est proche d'un score parfait avec un élément qui n'a pas été détécté (97/98=0,99). Le type « -er » a lui un score de 0,95, avec 3 expressions non détéctées. Cela donne un rappel moyen de 0,97.

Enfin, nous pouvons calculer le f-mesure de cet échantillon grâce à la précision moyenne (0,56), ainsi qu'au rappel moyen (0,97).

$$2* = \frac{0,55 * 0,97}{0,55 + 0,97} = 0,70 \tag{4.1}$$

Le résultat de ce calcul est de 0,70. Ce résultat est assez moyen, car la précision est médiocre. On remarque par ailleurs que la précision se rapproche de celle des fautes présentes dans le tableau 4.2 : 0,56 pour les deux méthodes. Cela veut dire que notre algorithme récolte beaucoup de bruits, surtout dans la catégorie « -é » . En revanche, le rappel qui est très haut nous montre que nous n'avons pas beaucoup manqué les termes que nous recherchions. Le tableau ci-dessous résume les calculs mentionnés dans cette partie.

	ex	**-é**	**-er**	**TOTAL**
Précision	1	0,24	0,56	**0,55**
Rappel	1	0,99	0,95	**0,97**
F-mesure	1	0,39	0,70	**0,70**

TABLE 4.4 – Rappel, Précision et F-mesure de la méthode n°2

La catégorie « ex » est celle qui est la plus précise et la plus performante. La catégorie « -é » a collecté énormément de bruit dans cet échantillon, et ne dépasse pas 0,5 de f-mesure. Nous allons voir avec plus de détails quels sont les termes qui ne sont pas pertinents et comment nous pourrions les éviter.

4.3.2 Analyses qualitatives de la seconde méthode

Contrairement à la première méthode, nous avons cette fois-ci réussi à trouver des pratiques prescriptives. Sur les 200 tweets, nous avons trouvé 3 commentaires que l'on peut nommer de prescriptif. Ce qui fait 1,5% d'expressions prescriptives dans cet échantillon. L'ensemble de ces pratiques se trouvent dans la catégorie « ex », ce qui augmente le taux d'expression prescriptive à 5% dans cette catégorie. Voici les trois phénomènes récoltés :

ex	malgré que	*Malgré que, franchement la blonde ! #LRDS*
ex	croive	*Communiqué pour mon Fan Club : inscription au KIKADI réalisée... On y croise les doigts les types & on y croive ! #RMCLive*
ex	croive	*#GGRMC il a dit qui croivent ? ! ! Putain il est temps qu'il parte en vacances @XXX ? ? ? ?*

TABLE 4.5 – Valeurs prescriptives de la méthode n°2

Les trois phénomènes présentent des pratiques prescriptives assez différentes. La recherche de ces pratiques est assez ardue, parce qu'elle demande parfois une part d'interprétation du message. Le premier exemple ne critique pas explicitement le terme « malgré que », mais le terme « la blonde » est percu comme une insulte qui réfère à un individu ayant utilisé le terme « malgré que ». Le second tweet n'est pas clair, il nécessite le contexte, mais le tweet semble sarcastique, même si on ne peut pas s'assurer de cela. Le dernier commentaire prescriptif est le plus explicite. Il émet clairement une aversion envers le terme « croivent », en conseillant à l'individu de partir en vacances.

La catégorie des expressions régulières semble donc être la plus performante pour trouver notre phénomène. Nous allons maintenant observer quelles sont les erreurs que les deux autres catégories ont extraites. Le tableau 4.3.2 réfère les catégories des termes non-pertinents.

Type	**-é**	**-er**	**TOTAL**
AUX	12	1	13
ETIQ	6	7	13
EXP	7	9	16
N-3	6	/	6
PA	38	/	38
?	5	/	5
TOTAL	**74**	**17**	**91**

TABLE 4.6 – Typologie des termes non pertinents de la méthode n°2

Sur les 88 termes non-pertinents, nous avons répertorié six catégories que nous allons décrypter. La première, « AUX » concerne les fautes d'orthographe sur les auxiliaires. Ces fautes vont empêcher Spacy d'annoter le mot comme un auxiliaire, le script va donc détecter le terme comme étant pertinent à extraire. L'exemple 17 nous illustre cela avec l'auxiliaire « être » qui est relié au pronom « je » élidé.

17. *@XXX Vraiment jsuis tro attaché au 2 joueurs, y a peut être une légère préférence pour lacazette mais elle est infime*

Le type « ETIQ » regroupe cette fois les verbes qui sont mal étiquetés par Spacy. On trouve tous les mots ayant reçu l'étiquette « verbe », bien qu'il n'en soient pas. C'est par ailleurs la catégorie la plus présente dans le type « -er », car beaucoup de noms en « -er » sont étiquetés en verbes, comme « super » dans l'exemple 18. La solution pour éviter ces étiquetages erronés serait de générer le script avec le pipeline de Spacy la plus performante, afin d'éviter de limiter les erreurs. Cependant, en gagnant en performance, nous perdons en vitesse de génération du script, qui est déjà conséquente.

18. *Je suis morte de rire c'est super moche ce qu'il lui a fait #LRDS url_path*

Le type « EXP » répertorie les expressions qui viennent perturber l'extraction. Par exemple « commencer a » contient un auxiliaire, mais qui ne concerne pas le verbe « écouter », qui a été extrait.

19. *Bon bah, on va commencer a* ***écouter*** *du freeze, le colors show était lourd*

Le type « N-3 » concerne la catégorie « -é », elle concerne les verbes qui possèdent bien un auxiliaire, mais qui se trouvent à l'antépénultième position par rapport au particpe passé (soit le troisième mot). Notre script ne prend que les deux mots précédents le verbe, ce qui explique pourquoi on retouve ce genre d'occurrences, comme pour l'exemple 20. Analyser plus profondément les termes avant le participe passé suffirait pour régler ce problème.

20. *par contre on a pas assez parlé d'Hamilton, c'est un putain d'avion de chasse, de P17 à P7 ? ? ? ? ? ? ? ? ?*

Le type « PA » concerne aussi les participes passés. Le signe signifie participe adjectival, pour définir les verbes qui sont en fait des adjectifs. On compte aussi dans cette catégorie les participes passés employés seuls, ces deux types sont similaires et représentent plus de la moitié des données non-pertinentes des participes passés (38/71=0,54). L'exemple 21 montre deux participes passés seuls sans auxiliaires.

21. *Djoko **disqualifié**, Giannis **blessé**. Bon bah je vais regarder #TheBoys*

Le dernier type « ? » répertorie les occurrences qui sont normalement correctes, mais qui sont tout de même présentes. Dans 22, « joué » à été sélectionné alors qu'il y a un auxiliaire juste avant. Il est possible que ce soit dû à une erreur d'étiquetage de Spacy qui ne considère pas ces auxiliaires comme tel.

22. *@XXX (Après c'est con tu as joué Butler qui mène 3-0 contre les Bucks et Boston qui mène 2-1 contre Toronto)*

4.3.3 Conclusion de la seconde methode

La seconde méthode d'extraction nous a permis de récolter nos premières expressions prescriptives. Nous avons pu observer que la catégorie « ex » est la plus précise pour chercher des pratiques prescriptives. Les occurrences de « -é » sont les moins pertinentes et les moins précises. La cause principale vient des participes passés sans auxiliaires, très présents dans cet échantillon.

Au final, les commentaires concernant les erreurs de conjugaison entre le participe passé et l'infinitif n'ont pas autant de valeur prescriptive que nous le

pensions. Le ciblage des termes est plus efficace que la recherche quantitative des deux autres expressions, même si elle a été réduite, en comparaison à la première méthode. Nous allons voir si cela est toujours le cas dans notre dernière analyse.

4.4 Analyse de l'extraction des expressions prescriptives

4.4.1 Analyses quantitatives de la troisième méthode

Après avoir appliqué le script au corpus, nous avons généré un tableau de 600 phénomènes, soit 33 fois moins que la méthode précédente (20 292) sur le même nombre de données. Nous allons voir la répartition des données. Nous analyserons toujours un échantillon de 200 occurrences, extrait de manière proportionnelle à la quantité des expressions dans ce tableau. Voici la répartition des cinq expressions :

Type	**quantité**	**part**	**%**
langage	49	16	8%
on dit	197	66	33%
orthographe	42	14	7%
le terme	65	22	11%
le mot	247	82	41%
TOTAL	**600**	**200**	**100%**

TABLE 4.7 – Répartion des termes prescriptives

On remarque tout de suite que les expressions « on dit » (33%) et « le mot » (41%) représentent à eux seuls 74% du document. Les 26% restants sont « langage » (8%), « orthographe » (7%) et « le terme » (11%).

Les mesures de précision et de rappel vont nous aider à savoir si ces données sont pertinentes. Sur les 200 termes, seuls 25 sont considérés comme étant non-pertinents. La précision est donc de 0,88 (1-(25/200)), ce qui est un très bon score. Le rappel montre que le script est passé à côté d'un terme pertinent, ce

qui donne un rappel presque parfait de 0,99 (175/176).

$$2* = \frac{0,88 * 0,99}{0,88 + 0,99} = 0,93 \tag{4.2}$$

Le score de f-mesure est excellent avec 0,93. Ce score est de 22 points supérieurs à celui de la méthode précédente (0,71), ce qui nous montre l'efficacité de cette extraction et nous prouve que l'extraction ciblée de termes est plus efficace. Cependant, notre objectif est d'abord de trouver des expressions prescriptives. Nous allons donc regarder en détails si des termes prescriptifs ont été extraits.

4.4.2 Analyses qualitatives de la troisième méthode

Cette méthode est la plus riche en expressions prescriptives. Nous avons en tout 36 pratiques prescriptives, ce qui est 12 fois supérieur à la méthode précédente (3). Cela représente 18% de l'échantillon. Si l'on regarde les catégories en détails, « on dit » est celle qui a le plus de pratiques prescriptives (17), soit 26% du total de cette catégorie. C'est « orthographe » qui a le taux de pratiques prescriptives le plus haut avec 29%. Au contraire, l'expression « le mot » possède 7 pratiques prescriptives, mais qui représentent 8,5% du total de la catégorie.

La colonne « meta » désigne les expressions qui sont métalinguistiques, mais qui n'ont pas de valeur prescriptive. Comme nous le pensions, la catégorie « le mot » produit beaucoup de métadiscours (51%). Voici un exemple de métadiscours non prescriptif :

23. *@XXX @XXX Alors utilise le mot « apprendre » si tu préfères.*

La catégorie « ? »liste les occurrences qui n'ont pas la certitude d'être prescriptifs, cela peut arriver quand le contexte n'est pas nécessaire pour comprendre le tweet. L'exemple 24 illustre cela.

24. *Même Léna et moi on dit pas ça url_path*

La catégorie « non » désigne les occurrences qui ne sont pas métalinguistiques, elles représentent presque la moitié de l'échantillon (47%). Nous avons pu analyser grâce à la précision que 25 occurrences ne sont pas pertinentes. Il

s'avère que toutes les occurrences sont présentes dans la catégorie « le mot ». L'erreur vient d'une imprécision dans l'expression régulière, elle prend les caractères autour du mot, ce qui peut extraire des mots qui contiennent les mêmes caractères. Une modification de l'expression régulière suffit pour régler ce souci. Voici un exemple de cette anomalie :

25. *Sainz il va faire chauffer le moteur d'Hamilton*

La catégorie « on dit » possède 37 « non », qui est la plus importante. Un grand nombre de ses expressions n'est pas prescriptif, car ce terme est utilisé dans beaucoup d'autres contextes sans métadiscours. Le terme est utilisé de manière déclarative, et non sous forme d'ordre à l'impératif, dans l'exemple 26.

26. *RT @XXX : Qiand on dit cheveux long c'est à la Leonardo DiCaprio des 90's pas djadja et dinaz*

	?	**meta**	**non**	**pres**	**TOTAL**
langage	1	4	7	3	15
on dit	1	11	37	17	66
orthographe	/	2	8	4	14
le terme	1	7	9	5	22
le mot	2	42	31	7	82
TOTAL	**5**	**66**	**93**	**36**	**200**

TABLE 4.8 – Répartion des pratiques prescriptives

Parmi les 36 pratiques prescriptives, on trouve des termes très variés. On trouve tout de même 11 occurrences des termes « yeti » et « mister freeze » venant d'un débat d'appellation d'une glace. Nous trouvons des termes régionaux (chocolatine). Certains tweets ne sont pas prescriptifs envers un mot, mais envers l'utilisation du langage en général, comme par exemple :

27. *il fait pas autant pitié que ton orthographe*

4.4.3 Conclusion de la troisième méthode

Les résultats de cette méthode sont très concluants. La recherche d'expressions est plus performante que la recherche de fautes. La méthode de recherche

est la plus simple des trois, car elle vise des termes en particulier en évitant de généraliser les pratiques prescriptives. Il serait assez simple d'améliorer la recherche et de supprimer du bruit en corrigeant le défaut de récolte de l'expression « le mot ».

4.5 Conclusion générale des analyses

L'analyse des trois méthodes montre clairement que la troisième méthode est de loin la plus performante pour récolter des pratiques prescriptives. La première méthode n'est pas pertinente, car les tweets sont un type de données textuelles qui ne sont pas soumises à un jugement des utilisateurs en cas de transgression de la norme. Par ailleurs, c'est la seule méthode qui n'a pas récolté de pratiques prescriptives. Cette méthode serait plus intéressante sur des données accordant plus d'importance à la norme. La seconde méthode est plus performante, car elle récolte moins de bruit et arrive cette fois à récolter des pratiques prescriptives. Les occurrences concernant les participes passés ont été moins intéressants que le ciblage d'expressions à tendance prescriptive.

La dernière méthode s'intéresse directement à notre phénomène en omettant la phase de recherche des fautes. Nous avons aussi ciblé des termes d'expressions prescriptives afin de maximiser nos chances de trouver les pratiques. C'est en récoltant le moins de phénomènes que nous avons eus le plus de pratiques, ce qui démontre de la qualité de cette méthode.

Chapitre 5

Conclusion

Lors de ce mémoire, nous avons émis la problématique suivante : **comment peut-on identifier et extraire les pratiques prescriptives des non-experts de la langue sur les réseaux sociaux ?** Nous y avons répondu en testant trois méthodes sur un corpus de tweets, tout en affinant les critères de séléction à chaque méthode. On peut dire qu'identifier notre phénomène est une tâche qui demande un certain degré de granularité. Les résultats nous ont montré qu'il n'y a pas de solutions pour extraire toutes ces pratiques, mais qu'en se penchant sur des cas particuliers, nous sommes capables d'en extraire. Cela affirme notre hypothèse disant que la recherche de ces termes serait complexe.

La seconde hypothèse s'attarde sur notre corpus, plus précisément sur le fait que les réseaux sociaux sont un lieu propice aux pratiques prescriptives. Les réseaux sociaux possèdent l'avantage d'être très flexibles vis-à-vis de la norme de la langue en général, ce qui attire les pratiques prescriptives. On peut affirmer cette hypothèse dans le cas où nous y avons extrait des pratiques prescriptives. Nous pouvons cependant infirmer cela pour la première méthode quantitative, qui n'est pas compatible avec des données aussi peu standardisées que les nôtres.

Ce travail nous montre surtout où placer le curseur dans l'automatisation pour optimiser les résultats. Nous avons pu être confrontés aux limites d'automatiser un phénomène aussi complexe que les pratiques prescriptives. La dimension humaine est encore essentielle dans l'extraction de ces termes. Nous avons

observé des soupçons d'ironie, de sarcasme, mais aussi des tournures de phrase implicites qui donnent du fil à retordre aux machines. L'outil reste essentiel au tri des données, mais la dimension implicite l'empêchera de prendre la décision d'extraire les données pertinentes.

Même si nous avons réussi à extraire des pratiques prescriptives, nous n'avons pas la certitude de l'authenticité des non-experts sur les réseaux sociaux. Il est impossible de vérifier si tous les internautes sont des non-experts. En dehors des réseaux sociaux, nous pouvons changer la modalité des données et se pencher sur des données orales. L'avantage est qu'en construisant un corpus oral, nous pouvons ajouter dans les métadonnées l'expertise de la langue que possèdent les locuteurs, et donc certifier la non-expertise des locuteurs. Nous pouvons donc approfondir cette recherche en l'axant sur des données orales.

Bibliographie

(1694). *Le dictionnaire de l'Académie françoise, dédié au Roy. T. 1. A-L.* Vve J. B. Coignard et J. B. CoignardVve J. B. Coignard et J. B. Coignard.

Abbou, J. (2022). Tenir sa langue. *GENRE !*

Amadieu, J.-B. (2014). Quand les écrivains canoniques écrivent mal. (2). Number : 2 Publisher : ITEM, Institut des textes et manuscrits modernes, UMR 8132 CNRS/ENS.

Andrews, L. (2006). Language exploration and awareness : a resource book for teachers.

Arana et Burnett (2023). Mathematical hygiene. *Synthese*.

Bird, S., Klein, E. et Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc.

Bush, V. (1945). As we may think.

Calabrese, L. et Rosier, L. (2015). Les internautes font la police : purisme langagier et surveillance du discours d'information en contexte numérique. (2):120–137.

Cameron, D. (1995). Verbal hygiene.

Damar, M.- (2010). De la polymorphie du purisme linguistique sur l'internet. 131(1):113–130. Place : Paris Publisher : Éditions de la Maison des sciences de l'homme.

De Clercq, K., Haegeman, L., Lohndal, T. et Meklenborg, C. (2023). Adverbial resumption in verb second languages. *Oxford Studies in Comparative Syntax*.

Delvin, J., Chang, M.-W., Lee, K. et Toutanova, K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding.

Donati et Graffi (2022). La grammatica generativa.

Dutant, J. (2012). *Grand dictionnaire de la philosophie*, page 863. Larousse.

Grishman, R. (1997). Information extraction : Techniques and challenges.

Hare et Mervyn (1952). *The Language of Morals*. Oxford University Press.

Hochreiter et Schmidhuber (1997). Long short-term memory.

Honnibal, M. et Montani, I. (2017). spaCy 2 : Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Humphries, E. (2023). Croient or croivent : French language commentary on twitter. 62(3):314–333. Publisher : Edinburgh University Press.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. et Stoyanov, V. (2019). Roberta : A robustly optimized bert pretraining approach.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É. V., Seddah, D. et Sagot, B. (2020). Camembert : a tasty french language model. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Mekki, J. (2023). Tremolo_tweets_corpus. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

Moïse, C. (2015). Lol non tkt on ta pas oublié. (167). Number : 167-168 Publisher : Association CRESEF.

New, B., Pallier, C., Brysbaert, M. et Ferrand, L. (2004). Lexique 2 : A new french lexical database. *Behavior Research Methods, Instruments, amp ; Computers*, 36(3):516–524.

Osthus, D. (2018). À la recherche du « locuteur ordinaire » : vers une catégorisation des métadiscours. (14):18–32. ISBN : 9782379060014 Number : 14 Publisher : Presses Sorbonne Nouvelle.

Robertson et Jones (1976). Relevance weighting of search terms. pages 112–124.

Salton, Wong et Yang (1975). A vector space model for automatic indexing. pages 613–620.

Tavosanis, M. (2007). A causal classification of orthography errors in web texts.

Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser et Polosukhin (2017). Attention is all you need.