Institut National des Langues et Civilisations Orientales

Département Textes, Informatique, Multilinguisme

Caractérisation objective des catégories textuelles pour le TAL : classification non-supervisée basée sur des descripteurs linguistiques

MASTER

TRAITEMENT AUTOMATIQUE DES LANGUES

Parcours :

Ingénierie Multilingue

par

Marina SEGHIER

Directeur de mémoire :

Patrick Paroubek

Encadrants :

Alice Millour et Jean-Yves Antoine

Année universitaire 2022/2023

Remerciements

Je tiens à remercier chaleureusement mes encadrants, Alice Millour et Jean-Yves Antoine, qui m'ont accompagnée tout au long de ce stage de recherche. Merci pour leur supervision éclairée, leurs précieux conseils et leur confiance. Grâce à eux, j'ai pu mettre mes compétences au service d'un sujet passionnant, progresser en rédaction scientifique et alimenter ma réflexion. Je suis très heureuse et fière de pouvoir continuer à travailler à leurs côtés en thèse, ces trois prochaines années.

Je tiens également à remercier Yoann Dupont pour son aide et son introduction à SemTagger, mon encadrant d'université Patrick Paroubek pour son suivi et ses conseils, ainsi que les membres de l'équipe PASTIS du laboratoire LIASD pour leur accueil et leur bienveillance.

Je remercie également Liam Duignan et Kevin Quach, mes acolytes de labo, sans qui l'été aurait été bien différent.

Table des matières

Table des figures

Liste des tableaux

Introduction

« *Je dis souvent que lorsque vous pouvez mesurer ce dont vous parlez, et l'exprimer en chiffres, vous en savez quelque chose ; mais lorsque vous ne pouvez pas le mesurer, lorsque vous ne pouvez pas l'exprimer en chiffres, votre connaissance est d'un genre maigre et insatisfaisant ; c'est peut-être le début de la connaissance, mais vous avez à peine, dans votre pensée, avancé jusqu'à l'état de Science, quelle que soit la question.* » Lord Kelvin (1824–1907)[1]

Ces paroles énoncées par le physicien en 1883 dans son ouvrage « *Popular Lectures and Addresses* » n'ont jamais eu autant de résonance que ces dernières années dans le domaine du Traitement Automatique des Langues (TAL).

En effet, les systèmes développés aujourd'hui sont de plus en plus répandus, utilisés, et présentés comme étant très performant pour un grand nombre de tâches. Par exemple, CamemBERT[2] obtient une F-mesure de 94,20 % pour la tâche d'annotation en dépendances syntaxiques, Stanza[3] (anciennement StanfordNLP) obtient une F-mesure de 96,07 % pour la tâche d'étiquetage morphosyntaxique sur du vieux français[4]), ou encore SpaCy[5] obtient une précision de 100 % pour la tâche de tokenisation du français.

Or, dans les faits, les performances annoncées ne sont pas toujours celles rencontrées selon les types de ressources textuelles auxquels ces outils sont confrontés. Leur comportement est fortement influencé par leurs données d'apprentissage et même s'ils ont une certaine capacité de généralisation, un système qui aura été entraîné sur du contenu encyclopédique comme *Wiki-*

1. Original : « *I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it ; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind ; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the state of Science, whatever the matter may be.* »
2. https://huggingface.co/Jean-Baptiste/camembert-ner
3. https://stanfordnlp.github.io/stanfordnlp/performance.html
4. Corpus : https://github.com/UniversalDependencies/UD_Old_French-SRCMF
5. https://spacy.io/models/fr#fr_core_news_sm

pédia ou du contenu journalistique comme *l'Est Républicain* par exemple, n'aura pas les mêmes performances sur d'autres types de texte, comme de la parole transcrite, ou encore de la poésie.

C'est ce qu'ont observé Millour et al. (2022) avec FENEC, un corpus d'évaluation réunissant six « genres » textuels (encyclopédie, poésie, prose, journalistique, parole transcrite et multi-sources), réalisé dans l'optique d'évaluer et de comparer trois outils d'annotation automatique (SpaCy, CasEN et Flair) *a priori* performants, sur la tâche de Reconnaissance d'Entités Nommées (REN) de plusieurs types (de lieux, de personnes, d'organisation, concepts nominaux, etc). Cette étude a permis d'éprouver la robustesse de ces outils et d'observer un différentiel de performances important entre les différents types de texte pour la tâche de REN. Cependant, elle présente aussi des limites car elle se fonde sur une classification des genres qui n'est pas caractérisée linguistiquement, et n'explique finalement pas ce différentiel.

L'apprentissage par transfert (*transfer learning*) ou peaufinage/raffinement (*fine-tuning*) est un moyen utilisé aujourd'hui de pallier ces chutes de performances et rendre un modèle plus performant sur des données cibles qui lui sont inconnues ; mais cette adaptation se fait bien souvent « à l'aveugle », sans étude des caractéristiques linguistiques qui mettent en difficulté les systèmes.

Nous abordons ici la question de la variation à travers celle des catégories textuelles, telles que le *genre*, les *domaines*, les *registres* et les *modalités*, des concepts qui sont parfois difficiles à définir et à délimiter. Nous verrons qu'un même terme peut recevoir des définitions différentes, dans des domaines liés ou apparentés. Mais aussi que ces notions forment un continuum : elles sont utiles, mais difficile à mobiliser pour une catégorisation stricte. En effet, nous verrons que la catégorisation n'est pas toujours faisable et qu'il est possible de trouver plusieurs « types » au sein d'un seul et même texte. Par exemple, un roman (genre) policier (domaine) ou encore du théâtre (genre) lyrique (registre).

Face aux variations observables dans la langue, toutes ces typologies sont autant de tentatives de classer les textes. Cependant, les catégories textuelles traditionnelles (genres/domaines/registres/modalités) définies *a priori* sans justification linguistique sont-elles pertinentes/optimales dans le cadre du développement et de l'adaptation d'outils pour la tâche de reconnaissance des entités nommées ? En effet, celles-ci n'*expliquent* pas la variation textuelle. Or, la caractériser nous permettrait de mieux prédire la performance d'un outil sur un texte de type « inconnu », et de concevoir des applications en TAL robustes à la variation des données.

Afin de répondre à cette question, nous nous proposons premièrement d'identifier, et deuxièmement d'analyser des descripteurs linguistiques perti-

nents pour la classification textuelle dans le cadre de la tâche de REN en français. Après une sélection des caractéristiques les plus pertinentes pour notre étude, nous procédons à une Analyse en Composantes Principales (ACP), une technique statistique exploratoire qui consiste à résumer l'information contenue dans un grand jeu de données en un certain nombre de variables synthétiques, combinaisons linéaires des variables originelles.

Le premier chapitre de ce mémoire sera consacré à un état de l'art des différentes typologies textuelles, puis des recherches dans le domaine de la catégorisation textuelle, ainsi que des travaux actuels sur l'observation de traits linguistiques pour une évaluation plus interprétable. À la suite de quoi dans le deuxième chapitre, nous présenterons la méthodologie suivie, de la préparation des données jusqu'à l'analyse en composantes principales. Enfin dans le troisième chapitre, nous mettrons en lumière les résultats obtenus et discuterons des perspectives.

Chapitre 1

Etat de l'art

Dans ce chapitre consacré à l'état de l'art, nous tenterons de définir dans un premier temps, certaines typologies textuelles existantes, notamment les genres, domaines, modalités et registres ; puis dans un deuxième temps, nous présenterons les travaux qui sont à l'origine de cette étude, Biber (1988) et Passonneau et al. (2014), dont l'objectif était de caractériser le genre en identifiant des descripteurs linguistiques et de représenter la variation textuelle à travers de nouvelles dimensions (une nouvelle typologie textuelle) ; enfin dans un troisième temps, nous nous pencherons sur les travaux actuels qui mêlent observations des caractéristiques linguistiques et interprétation d'une évaluation.

1.1 Des typologies textuelles pour classer les textes

D'après Bidaud and Megherbi (2005), les différences entre la langue orale et la langue écrite se situent sur un continuum de pratiques langagières. En effet, nous pouvons parler et écrire dans des situations qui peuvent différer ou non, des conditions habituelles de réalisation ; par exemple, au cours d'activités à dominante formelle contrainte (lettre administrative, texte rédigé, discours politique, télévisé, etc) ou d'activités à dominante informelle (lettre à un ami, conversation classique, etc). Ainsi, la production langagière quelle qu'elle soit, s'inscrit dans une situation donnée : tout comme à l'oral, un texte écrit est aussi un texte en situation.

Nous allons à présent tenter de définir les différentes typologies, selon lesquelles les catégories communément utilisées pour classer les différents types de textes découlent.

Le genre

Le Larousse défini le genre comme une « Catégorie d'œuvres littéraires ou artistiques définie par un ensemble de règles et de caractères communs ; style, ton d'un ouvrage ». Initialement littéraire et philologique, le concept de *genre* est par la suite devenu très utilisé dans le domaine de la Recherche d'Information (RI) et en catégorisation textuelle (Poudat et al., 2006). Un genre présente des propriétés linguistiques formelles qui permettent de l'identifier et de le différencier des autres ; par exemple, certains marqueurs du théâtre comme les points d'exclamations seront absents dans des textes encyclopédiques. On peut trouver plusieurs typologies textuelles des genres ; par exemple ceux de la comédie, tragédie, drame, roman, nouvelle, conte... chez Malrieu and Rastier (2001), mais aussi plus communément les genres narratif, poétique, théâtral, épistolaire, argumentatif, tel qu'on nous l'enseigne en cours de français aujourd'hui à l'école.

Le domaine

Contrairement aux genres, les *domaines* quant à eux se définissent plus sur le plan lexical que sur celui de la morphosyntaxe, dans la mesure où ils reflètent supposément un champ de connaissance spécifique (juridique, médecine, physique, informatique...) (Poudat et al., 2006). En effet, dans différentes techniques de classification par domaine (ou thème) de documents, les textes sont généralement réduits à l'état de sacs de mots et sont décrits par le vocabulaire qu'ils contiennent. Les mesures les plus fréquentes en classification thématique sont calculées sur les mots, leurs racines, les *clusters* de mots (regroupements de mots par « thèmes », « sujets », « topics »...) ou encore les champs lexicaux. Les domaines sont ainsi souvent décrits en termes de relations lexicales (ex : champs lexicaux de la maladie, de la santé, racines grecques et latines dans le domaine médical, etc), sans autre niveau d'analyse que lexical.

La modalité

Selon Bidaud and Megherbi (2005), les différences linguistiques entre l'oral et l'écrit peuvent être envisagées sur trois niveaux : le discours, la phrase et le mot.

Sur le plan de la phrase, une des différences majeures se situe dans le respect ou non de l'ordre des mots (sujet, verbe, objet). À l'oral, même s'il obéit à une logique, l'ordre des mots est plus libre. Par exemple, certains phénomènes linguistiques comme la dislocation (ex : « mon chien, *il* est blanc et noir ») et l'absence de la négation « ne » sont réservées à l'oral. En effet,

l'absence du « ne » serait plus souvent notable dans des conversations libres, que dans des discours préparés (politiques et juridiques par exemple).

Sur le plan du mot, même si la richesse du vocabulaire à l'écrit et à l'oral paraît *a priori* la même d'un individu à un autre, la fréquence d'usage des mots à l'écrit est inférieure à celle de l'oral. De fait, nous utilisons plus de mots rares à l'écrit qu'à l'oral sans doute pour des raisons de style, mais aussi parce que l'écrit n'est pas soumis aux mêmes contraintes temporelles que l'oral. Cette contrainte plus forte nous pousse à utiliser des mots qui sont plus facilement récupérables et par conséquent, un vocabulaire moins riche. Elle nous pousse également à produire plus de disfluences telles que les hésitations vocaliques, les marqueurs discursifs, les auto-corrections, les amorces, les répétitions, les tics de langage, les faux-départs, etc. Ces phénomènes font ainsi les spécificités de la parole transcrite : la segmentation des phrases est plus compliquée, les verbes ne sont pas aux mêmes temps...

Tout comme l'absence du « ne », ces phénomènes restent cependant plus fréquents dans le cadre de la parole spontanée, que de la parole préparée.

Le registre

Le terme *registre* (parfois considéré comme un synonyme de *genre*) a été employé pour la première fois par le linguiste Thomas Bertram Wallace Reid (Reid, 1956) pour désigner une variété linguistique appropriée à une situation d'élocution et intègre ainsi pour la première fois, une dimension sociale que l'on ne retrouve pas dans une autre typologie. Elle correspond aux *niveaux de langue* qui sont les registres courant, familier, soutenu. Mais nous retrouvons également la notion de registres littéraires, enseignée aussi à l'école, qui sont les registres comique, lyrique, tragique, satirique, épique, fantastique. La première est liée à la variation des situations d'énonciation ; la deuxième est liée au ton, à l'effet particulier que produit le texte sur le lecteur.

Donner une définition fixe de ce qu'est un type de texte s'avère être une tâche difficile. Nous pouvons voir qu'il existe différentes classifications qui reposent sur différents niveaux : d'après ces définitions, nous pourrions comparer le genre à la « forme », le domaine au « contenu », le registre au « ton » et la modalité au « mode ». Mais ces classifications ne sont pas suffisantes : sans critères formels de caractérisation, elles présentent des limites. C'est pourquoi tout au long de ce mémoire, nous préférerons le terme « catégorie » au terme « genre ».

La prochaine section présente l'une des premières tentatives de définir la variation textuelle au travers de plusieurs dimensions linguistiquement fondées, réalisée et décrite par Douglas Biber dans son ouvrage « *Variation across speech and writing* » (1988).

1.2 Précédents dans la recherche sur la classification textuelle

1.2.1 Émergence de nouvelles dimensions en anglais britannique

Biber (1988) a été le premier à mener une étude statistique à grande échelle sur ce qu'il appelle « genre », dans le but d'identifier plusieurs dimensions de la variation dans la langue. Il a procédé à une Analyse en Composantes Principales (ACP) sur deux corpus de l'anglais britannique, auxquels il a ajouté une collection de ses propres lettres manuscrites professionnelles et personnelles.

Le premier corpus est le *Lancaster-Oslo-Bergen* (LOB) et compte près d'un million de mots. Il couvre la période des années 1970 et est composé de plusieurs échantillons de « genres » suivants : presse (reportage, éditorial, critiques), religion, « compétences et loisirs », « tradition populaire », biographies, documents officiels, prose académique, fiction (générale, mystère, science-fiction, aventure, romantique), humour. Nous pouvons voir que D. Biber considère ces différents types de texte comme des genres, mais certains se définiraient aussi par les notions de registres, domaines ou modalités.

Le deuxième est le *London-Lund* et compte approximativement 500 000 mots. Il a été compilé des années 1970 à 1990, et est composé de 87 textes de parole transcrite, issus de conversations privées, publiques, téléphoniques, d'émissions, de discours spontanés et préparés.

Ses travaux reposaient sur 67 caractéristiques, extraites à l'aide de méthodes *ad hoc*, dont on ne connaît la fiabilité. En effet, D. Biber traite sans trop entrer dans les détails, d'un programme de *tagging* qu'il a utilisé qui consistait en plusieurs listes de mots dans un dictionnaire, et fonctionnait en deux étapes : (1) identification initiale de la catégorie grammaticale de chaque mot, et (2) dans les cas où le dictionnaire répertoriait plus d'une catégorie grammaticale possible pour un mot, résolution d'ambiguïtés par une succession de tests *if* (ex : si le mot finit par *-ing*, alors c'est un participe présent ; si le mot finit par *-ly*, alors c'est un adverbe, etc).
Voici un extrait des listes fermées des catégories grammaticales :

DO : do, does, did, don't, doesn't, didn't, doing, done
HAVE : have, has, had, having, -'ve#, -'d#, haven't, hasn't, hadn't
Possessive pronouns : my, our, your, his, their, its (plus contracted forms)
Demonstratives : this, that, these, those
Quantifiers : each, all, every, many, much, few, several, some, any
Numerals : one ... twenty, hundred, thousand

Ces caractéristiques étaient la fréquence d'occurrence d'éléments lexicaux (ex : amplificateurs comme *absolutely*, *altogether*, *completely...*), des parties du discours et des caractéristiques quasi-syntaxiques (telles que la coordination, la négation, l'effacement des pronoms relatifs, les *that-clauses*[1] etc).

Au terme de son « analyse factorielle » (analyse en composantes principales) à sept facteurs, D. Biber a affirmé que la variation linguistique était continue selon six dimensions : 1) impliqué (affectivement) *VS* informationnel, 2) narratif *VS* non-narratif, 3) référence explicite *VS* dépendante de la situation, 4) expression manifeste de persuasion, 5) abstrait *VS* non-abstrait et 6) production d'informations sous contrainte temporelle.

Mais à la différence des genres, ces dimensions sont rarement mobilisées dans la conception d'outils de TAL pour décrire les disparités entre les contenus textuels.

Plus tard, Biber (1993) s'est demandé comment atteindre la « représentativité » dans la conception de corpus linguistiques (ce que signifie « représenter » une langue, la taille d'échantillons requise dans un corpus, etc). Il soutient qu'une étape préliminaire serait d'identifier les paramètres situationnels qui distinguent les textes d'une communauté linguistique et d'identifier l'éventail de caractéristiques linguistiques importantes qui seront analysées dans le corpus. Observer les propriétés linguistiques d'un texte s'avère alors indispensable, de sa conception à son usage en informatique, notamment pour une meilleure interprétabilité de la performance des outils avec lesquels on l'utilise.

La section suivante présente le travail de Passonneau et al. (2014) qui s'est appuyé sur celui de Biber (1988), mais diffère par le corpus et les caractéristiques choisies. L'objectif était de reprendre l'hypothèse selon laquelle la variation des genres est continue selon de multiples dimensions.

1. *That* est utilisé comme conjonction pour lier un verbe, un adjectif ou un nom à une proposition (*clause* en anglais.)

1.2.2 Retrouver ces dimensions en anglais américain

Passonneau et al. (2014) ont utilisé le corpus *Manually Annotated Sub-Corpus* (MASC) en anglais américain (extrait du *Open American National Corpus* (OANC)), dont la période s'étend des années 1990 à nos jours et compte près de 500 000 mots. Son contenu est très hétérogène puisqu'il inclue un large éventail de genres traditionnels, issus de l'écrit (documents gouvernementaux, techniques, journal, actualité, essai, fiction, non-fiction, lettre, guide de voyage, scénario de films, blagues) comme de l'oral (formel d'une part (transcriptions d'audiences, débats), informel d'autre part (un genre qui regroupe entretiens et conversations téléphoniques)), mais aussi de nouveaux médias sociaux de notre époque (e-mails, blogs, Twitter, spams) et de fanfictions (dites *ficlets*)).

Avant de faire une ACP pour tenter de retrouver les dimensions proposées par Biber (1988), ils ont procédé à un regroupement hiérarchique (*clustering*) non-supervisé de l'ensemble du jeu de données, en utilisant la distance de Manhattan (connue également sous le nom de taxi-distance), similaire à la distance euclidienne mais moins sensible aux valeurs aberrantes. On retrouve dans les *clusters* obtenus, plusieurs dimensions de la variation identifiées par D. Biber. En effet, l'expérience a identifié six groupes majeurs pour les 19 « genres » de MASC : deux singletons (guides de voyage et documents techniques), un groupe de trois genres de l'oral (formel : transcriptions d'audience, de débats ; et informel : les entretiens et conversations téléphoniques), deux groupes de quatre genres (l'un regroupe les genres « narratifs » : fiction, scénario de films, blagues, ficlets ; l'autre regroupe les genres « interactifs en ligne » (oral) et écrits : lettres, e-mails, spams, tweets), et enfin, un groupe de six genres (blog, essai, journal, actualités, non-fiction et documents gouvernementaux). Ce résultat apporte une certaine justification concernant à la fois la sélection des caractéristiques choisies, mais également les genres définis dans le corpus MASC.

Certaines des caractéristiques utilisées dans les travaux de Biber (1988) ont été reprises dans l'ACP, mais la plupart ont été choisies à partir du riche ensemble d'annotations validées manuellement qu'offre le corpus MASC, y compris celles produites par des outils tels que les étiqueteurs en parties du discours, les détecteurs d'entités nommées, et autres. D'abord au nombre de 37, la caractéristique des *mots étrangers*, des *interjections*, des *verbes de persuasion* (classe de verbes en anglais) et des *exclamations* ont été retirées de l'expérience car leurs valeurs étaient aberrantes. Ils ont donc réalisé l'ACP avec 33 caractéristiques.

Au terme de leur étude, ils valident l'hypothèse selon laquelle la variation des genres est continue dans de multiples dimensions en raison de propriétés

contextuelles comme des contraintes cognitives et l'interactivité.

1.3 La corrélation entre descripteurs linguistiques et évaluation des performances

Dans une étude visant à proposer une méthode d'évaluation généralisée, fine et multi-jeux de données pour la tâche de reconnaissance d'entités nommées, Fu et al. (2020) avancent qu'une simple mesure holistique telle que la précision ou la F-mesure ne nous permettent pas d'expliquer pourquoi ni comment des méthodes particulières (ex : un modèle à base de CRF) fonctionnent différemment d'un jeu de données à un autre.

Ils ont ainsi exploité la notion de caractéristiques linguistiques sous le terme d'« attributs ». Il s'agit de valeurs qui caractérisent les propriétés d'une entité nommée. Calculées à partir de ces dernières, elles seraient corrélées avec les performances d'un système de REN. Ces attributs peuvent être par exemple la longueur d'une entité nommée en tokens, la fréquence d'un token ou d'une entité, etc.

Dans la continuité de ce travail, Liu et al. (2021) présentent EXPLAINABOARD, un outil en ligne qui permet aux chercheurs de diagnostiquer les forces et les faiblesses d'un système (ex : en quoi le système le plus performant est-il mauvais ?), d'interpréter les relations entre plusieurs systèmes (ex : en quoi le système A surpasse le système B ?) et d'examiner de près les résultats des prédictions (ex : quelles sont les erreurs courantes commises par un système ?).

Là encore, les propriétés linguistiques ont une place prépondérante puisqu'elles permettent d'interpréter les performances des systèmes. En effet, une des tâches de cet outil est d'analyser des biais dans les données, en déterminant quelles sont les caractéristiques des différents jeux de données évalués.

Dans le cadre de cette étude, nous n'avons pas pu exploiter ces différentes pistes afin d'améliorer l'interprétabilité des évaluations faites, faute de temps, mais il serait intéressant de s'y pencher pour de futurs travaux.

Chapitre 2

Du corpus brut à l'analyse en composantes principales

L'objectif de ce travail a été de reproduire celui de Passonneau et al. (2014), adapté à la langue du français. La méthodologie décrite dans cette partie se divise en trois grands axes : dans un premier temps, nous présentons les données sur lesquelles nous avons travaillé, à savoir le corpus FENEC, puis dans un deuxième temps, nous détaillerons comment nous avons préparé nos données, de la tokenisation jusqu'aux calculs des caractéristiques (traduit de l'anglais *feature* ; correspond à une variable, une propriété – dans notre cas, linguistique – calculée à partir d'un jeu de données), avant d'exposer dans un troisième temps, la méthode de l'Analyse en Composantes Principales (ACP) et les différentes expériences menées.

2.1 Un corpus varié

FENEC (*FrEnch Named-entity Evaluation Corpus*) (Millour et al., 2022) est un corpus d'évaluation pour la tâche de reconnaissance d'entités nommées en français. Il est composé de onze documents de six genres textuels différents et annoté manuellement en entités nommées selon le schéma fin Quaero[1]. Les auteurs présentent une correspondance entre le schéma d'annotation Quaero et les jeux d'étiquettes de CasEN et de **WikiNER** (quatre étiquettes : LOC, PER, ORG, MISC) afin de pouvoir jouer sur différents grains d'annotation.

1. Guide d'annotation : http://www.quaero.org/media/files/bibliographie/quaero-guide-annotation-2011.pdf

Il a été élaboré dans le but d'entamer un travail comparatif sur l'interprétation de la performance de différents outils (CasEN[2], SpaCy[3] et Flair[4]) pour la tâche de reconnaissance des entités nommées, qui comprend deux parties : la segmentation et la classification (ou typage) des entités.

Dans le cadre de notre étude, le corpus nous permet d'explorer les catégories textuelles de la poésie (catégorie éponyme), de l'essai et du roman (catégorie « prose »), aux domaines encyclopédique (catégorie « encyclopédie ») et journalistique (catégorie « informations ») et dans le même temps la modalité écrite pour ces derniers, mais aussi orale (catégorie « parole »).

2.1.1 Catégorie *prose*

42131-0 : Ce document est un échantillon du *Traité sur la Tolérance* de Voltaire, datant du XVIIIe siècle. Il contient 40 phrases et 1 020 tokens, et est sous licence Project Gutenberg[5].

pg6470 : Le texte de ce fichier est un extrait du *Ventre de Paris* d'Émile Zola, datant du XIXe siècle. Il comptabilise 51 phrases et 1 002 tokens, et est également sous licence Project Gutenberg.

2.1.2 Catégorie *poésie*

pg6099 : Ce document est un échantillon des *Fleurs du Mal* de Baudelaire, datant du XIXe siècle. Il contient 30 phrases et 1 014 tokens, et est également sous licence Project Gutenberg.

56708-0 : Ce document est un échantillon d'*Œuvres d'Arthur Rimbaud - Vers et proses*, datant du XIXe siècle. Il contient 52 phrases et 1 027 tokens, et est également sous licence Project Gutenberg.

2.1.3 Catégorie *parole*

Spoken : Cet échantillon contient de la parole transcrite et est tiré de la ressource Rhapsodie Lacheret, Anne et al. (2014), un Treebank annoté en syntaxe et en prosodie pour l'analyse du discours en français parlé, datant de notre siècle. Il contient 70 phrases et 1 028 tokens, et est sous licence CC BY-SA 4.0.

2. https://tln.lifat.univ-tours.fr/version-francaise/ressources/casen
3. https://spacy.io/api/entityrecognizer
4. https://huggingface.co/flair/ner-french
5. https://www.gutenberg.org/policy/license.html

2.1.4 Catégorie *encyclopédie*

WikiNER : Cet échantillon (français) est extrait d'un corpus multilingue Nothman et al. (2013) (anglais, allemand, français, polonais, italien, espagnol, néerlandais, portugais, russe), issu du texte, de la structure des articles et des métadonnées de Wikipédia, y compris les titres, liens, catégories, modèles, boîtes d'informations et données de désambiguïsation. Il contient 36 phrases et 1 003 tokens, et est sous licence CC BY 4.0.

2.1.5 Catégorie *informations*

APIL : Cet échantillon est extrait d'un programme culturel produit par l'office du tourisme d'Othe-Armance et contient des informations touristiques. Il contient 29 phrases et 1 002 tokens, et est sous licence LGPL-LR.

Wikinews : Cet échantillon est composé d'articles issus du site Wikinews[6], une source d'informations qui se veut libre, neutre, vérifiable, collaborative et citoyenne. Il contient 46 phrases et 1 024 tokens, et est sous licence CC BY 2.5.

2.1.6 Catégorie *multi-sources*

UD French GSD : Cet échantillon est extrait du *treebank* du même nom, fait par *Universal Dependencies*, dont le contenu provient de blogs, d'informations, d'avis, et de Wikipédia. Il contient 35 phrases et 1 021 tokens, et est sous licence CC BY-SA 4.0.

Sequoia : Cet échantillon est extrait du corpus du même nom Candito and Seddah (2012), composé de textes provenant d'Europarl français, du journal l'Est Républicain, de Wikipédia Fr et des documents de l'Agence Européenne du Médicament. Il contient 44 phrases et 1 002 tokens, et est sous licence LGPL-LR.

French Question Bank : Cet échantillon est extrait du *treebank* Seddah and Candito (2016) du même nom, qui est composé de 2600 questions, dont les deux tiers sont le résultat d'un alignement avec l'*English Question Bank*. Il contient 102 phrases et 1 006 tokens, et est sous licence LGPL-LR.

En nous penchant plus en détail sur les documents du genre *multi-sources*, nous nous sommes rendus compte que les documents qui le composent étaient

6. `https://fr.wikinews.org/wiki/Accueil`

trop hétérogènes en terme de contenu. Nous avons donc pris la décision de basculer l'échantillon **UD French GSD** dans le genre *informations*, et d'exclure les échantillons de **Sequoia** et de **French Question Bank** (qui n'est pas vraiment un genre, mais un type de document) de l'expérience, ceci dans le but de ne plus avoir de genre *multi-sources* car peu interprétable.

Par rapport à l'expérience de Passonneau et al. (2014), bien que la distribution des genres au sein du corpus MASC était équitable (25 000 tokens environ par genre, soit 5-6% du corpus), la taille des textes au sein d'un même genre était différente. Ils ont alors concaténé tous les textes de chaque genre, puis procédé à un nettoyage des parties qui présentaient un excès de ponctuation ou de caractères spéciaux (comme dans les en-têtes de courrier électronique, les références bibliographiques ou le code informatique par exemple), afin de les diviser en échantillons de tailles égales, arrondis à la limite de la phrase la plus proche. Nous présentons dans la section suivante comment nous avons préparé les données.

2.2 Traitement des données

Passonneau et al. (2014) ont d'abord créé des échantillons de 1 000 tokens environ, en suivant l'observation faite par Biber (1993) qui est que les caractéristiques linguistiques (même les plus rares) sont relativement stables dans des échantillons de cette taille. Cependant, pour les caractéristiques utilisées dans leur étude, il s'est avéré que les résultats étaient similaires avec des échantillons de 500 tokens.

Étant donné la taille peu conséquente des 11 documents qui composent notre corpus (environ 1 000 tokens chacun), nous avons choisi de segmenter nos données en échantillons de 200 tokens. Pour cela, nous nous sommes penchés sur la façon dont nous voulions tokeniser nos données.

2.2.1 Le choix du segmenteur pour le corpus

La tokenisation est un procédé qui permet de transformer un texte en une série de tokens individuels, où chaque token représente un « mot ». Cependant, la notion de mot étant difficile à définir, nous emploierons ici le terme de « segment annotable ».

Pour les besoins de notre étude, nous voulions :

— garder les parties des différentes entités nommées qui sont séparées par des traits d'union ensemble (ex : *['ERVY-LE-CHATEL']* et non pas *['ERVY', '-', 'LE', '-', 'CHATEL']*), car si nous raisonnons en

partie du discours, l'entité entière est porteuse de l'information (ici, un nom propre) ;
— dans la mesure du possible, conserver les parties des autres segments annotables du corpus séparés par des traits d'union ensemble (ex : *['Là-haut']* au lieu de *['Là', '-', 'haut']*, et *['appelle', '-t-on']* au lieu de *['appelle', '-', 't', '-', 'on']*) car sur le plan morphosyntaxique, ils sont censés ne former qu'un seul et même « mot » (« *Là-haut* » est un adverbe, « *appelle* » un verbe et « *on* » un pronom, le « *t* » n'étant qu'une consonne euphonique qui a pour seule fonction d'assurer la liaison entre les deux) ;
— enfin, séparer les segments annotables grammaticaux (déterminants, prépositions, conjonctions, pronoms...) qui subissent l'élision, du mot qui les succède (ex : *[« l' », 'atmosphère']* et non pas *[« l'atmosphère »]*).

Nous avons alors essayé plusieurs tokenisers :

Tokeniser français de la librairie SpaCy. Le comportement du tokeniser sur les entités nommées avec des traits d'union était très variable selon les segments annotables, et ceci sans qu'on puisse l'expliquer. En effet, elle était correcte pour certaines :
— *['ERVY-LE-CHATEL']*, *['MARAYE-EN-OTHE']*, *['États-Unis']*, *['Tiangong-1']*...
... mais incorrecte pour d'autres, sans que l'on sache pourquoi :
— *['EAUX', '-', 'PUISEAUX']*, ***['état', '-', 'major']***, *['des', 'armées']*, *['Burkina', '-', 'Faso']*, *['Royaume', '-', 'Uni']*...

Fonction *word_tokenize* de la librairie NLTK (*nltk.tokenize*). La tokenisation des segments annotables grammaticaux élidés dépendaient de l'encodage de l'apostrophe (signe graphique de l'élision). En effet, elle était correcte si ce dernier était le caractère *guillemet-apostrophe* « ’ » (code Unicode : U+2019) :
— *['école', 'primaire', **'d', "'**, 'Estissac']*, *['Stage', 'être', 'soi', 'dans', **'l', "'**, 'univers']*... (document APIL)
... mais ne l'était pas s'il s'agissait du caractère *apostrophe* « ' » (code Unicode : U+27) :
— *['Peuples', 'de', « **l'**Orient »]*, *['Roi', « **d'**Assyrie »]*... (document 42131-0)

Segmenteur *SemTagger* du logiciel SEM[7], développé par Yoann

7. Site : https://www.lattice.cnrs.fr/sites/itellier/SEM.html, Github :

Dupont. La tokenisation fonctionnait très bien pour :

- les entités nommées : *['Pierre-le-Grand']*, *['EAUX-PUISEAUX']*, *['Burkina-Faso']*, *['troisième', 'président', 'des', 'États-Unis']*, *['état-major', 'des', 'armées']*...
- les segments annotables tels que : *['peut-être']*, *['vingt-neuf']*, *['a', '-t-il']*, *['appelle', '-t-on']* et même en écriture inclusive *['wikimédien.ne.s']*...
- les caractères *apostrophes* « ' » (code Unicode : U+27) : *['Peuples', 'de', « l' », 'Orient']*...
- les caractères *guillemet-apostrophes* « ’ » (code Unicode : U+2019) : *['Couvent', 'de', 'l'', 'Orée']*...

... mais était problématique pour des cas tels que :

- l'organisme « M5S » : ['M', '5', 'S']
- ou encore *['puante.–Pour']*, *['–démon–qui']*...

Ces derniers apparaissent dans le document *56708-0* (genre *poésie*), dont voici l'extrait :

Tu en es encore à la tentation d'Antoine. –l'herbe d'été bourdonnante et ***puante.–Pour*** *la fièvre des mères et des enfants. À Lulu,–****démon–qui*** *a conservé un goût pour les oratoires du temps des Amies et de son éducation incomplète.*

En voyant le premier cas, nous pourrions penser qu'ajouter des retours chariots au texte brut réglerait le problème, mais en voyant le deuxième, le caractère « – » n'est pas forcément synonyme de début de phrase. La ponctuation particulière de ce document met donc en difficulté ce segmenteur et semble être la cause de son comportement.

Nous pouvons constater qu'il est difficile de trouver un tokeniser qui réponde à tous nos critères. Mais à la suite de ces essais, nous avons décidé d'utiliser SemTagger lors d'une première étape (cf. section 2.2.2.) de caractérisation des entités nommées, puis de contrôler nous-même la segmentation des autres segments annotables à l'aide d'une expression régulière avec la fonction RegexpTokenizer de NLTK.

Expression régulière avec la fonction *RegexpTokenizer* de la librairie NLTK. Nous avons utilisé l'expression régulière suivante

`https://github.com/YoannDupont/SEM`

`\w'|\w+|[^\w\s]`, qui permet de tokeniser les segments annotables élidés avec leur apostrophe, et tout autre signe de ponctuation :

— « – *Toi, mon petit, je t'en.... et, tiens !* » devient *['-', '-', 'Toi', ',', 'mon', 'petit', ',', 'je', « t' », 'en', '.', '.', '.', '.', 'et', ',', 'tiens', '!']*

2.2.2 Caractérisation et délimitation des entités nommées

Le corpus FENEC ayant été annoté selon divers jeux d'étiquettes, nous avons choisi celui dont les catégories s'approchaient le plus de celles de Passonneau et al. (2014), soit le schéma d'annotation du corpus WikiNER[8], avec le jeu d'étiquettes suivant : LOC, PER, ORG, MISC (respectivement : *location* pour les noms de lieux (ex : Fukushima, Saturne, Burkina-Faso...), *person* pour les noms de personnes (ex : Hugues Mingarelli, Narcisse Pigeon, Horace...), *organization* pour les noms d'organismes (ex : Metro-Goldwyn-Mayer, Fonds des Nations unies pour l'enfance, lycée Arago...), *miscellaneous* pour le reste (« divers », ex : boulanger, Chroniques de l'oiseau à ressort, *Practicing Philosophy*...)). Nous avons travaillé à partir des textes bruts des documents du corpus et des fichiers d'annotations correspondants (extrait en figure 2.1).

```
T1  ORG 0 5 Valve
T5  ORG 88 91   M5S
T7  LOC 434 442 Belgique
T9  ORG 483 500 justice espagnole
T10 MISC 519 531    juges belges
T12 MISC 681 684    CRS
T14 ORG 756 767 Royaume-Uni
T16 MISC 786 803    diplomates russes
T18 ORG 844 852 Facebook
T19 ORG 855 861 Google
T20 MISC 912 919    AdWords
T21 PER 957 1004    Les forces de défense et de sécurité burkinabés
T24 ORG 1064 1069   Valve
T26 LOC 1174 1181   Algérie
T27 LOC 1184 1190   Égypte
```

FIGURE 2.1 – Extrait du fichier *wikinews-2018-03.sample.ann*.

Calcul du nombre de segments avec SemTagger

Une entité nommée peut être composée d'un ou plusieurs segments annotables. À cette étape, il ne s'agit donc pas simplement de découper le texte

8. https://metatext.io/datasets/wikiner

tous les 200 tokens, au risque de « casser » malencontreusement une entité nommée et d'avoir deux morceaux d'une même entité dans deux échantillons différents. Pour contourner ce problème, il était alors important de prendre en compte le nombre de tokens de chaque entité nommée, et de les considérer comme un seul et même bloc lors de l'échantillonnage.

Ainsi, nous avons calculé le nombre de segments annotables de chaque entité nommée et ajouté aux fichiers d'annotations (extrait en figure 2.2).

```
T1  ORG 0   5   Valve   1
T5  ORG 88  91  M5S 3
T7  LOC 434 442 Belgique    1
T9  ORG 483 500 justice espagnole   2
T10 MISC    519 531 juges belges    2
T12 MISC    681 684 CRS 1
T14 ORG 756 767 Royaume-Uni 1
T16 MISC    786 803 diplomates russes   2
T18 ORG 844 852 Facebook    1
T19 ORG 855 861 Google  1
T20 MISC    912 919 AdWords 1
T21 PER 957 1004    Les forces de défense et de sécurité burkinabés 8
T24 ORG 1064    1069    Valve   1
T26 LOC 1174    1181    Algérie 1
T27 LOC 1184    1190    Égypte  1
```

FIGURE 2.2 – Extrait du fichier *wikinews-2018-03.sample_ tab.ann*.

Création de fichiers intermédiaires

Nous avons ensuite créé des nouveaux fichiers dans lesquels les entités nommées ont été remplacées par leur code (ex : T1, T2, T3...). En figure 2.3, un extrait du fichier texte d'origine, et en figure 2.4, un extrait du fichier texte créé, version encodée.

```
Valve : de nouveaux jeux annoncés après plusieurs années sans sorties  9 mars 2018 .
Le M5S se présente généralement comme anti-système et œuvre pour la mise en place d'un
revenu universel , la baisse de l'impôt sur le revenu, la couverture des coûts de garde
des enfants et la mise en place de traités bilatéraux pour rapatrier les clandestins.
Le mandat d'arrêt européen lancé à son encontre le 2 novembre 2017 alors qu'il était en
Belgique, avait été retiré le 5 décembre, car la justice espagnole craignait que les
juges belges refusent de retenir les chefs d'inculpation, affaiblissant le dossier.
En 1952, il fonde sa maison de couture.
Puis, il s'est agressé à un groupe de CRS qui étaient en train de terminer leur footing.
La semaine dernière, le Royaume-Uni a déjà expulsé 23 diplomates russes.
— Quelques mois après le réseau social Facebook , Google renforce le contrôle des
publicités diffusées sur AdWords , son service de régie publicitaire.
Les forces de défense et de sécurité burkinabés ont riposté, abattant les 8 assaillants
et en capturant 2.
```

FIGURE 2.3 – Extrait du fichier *wikinews-2018-03.sample.txt*.

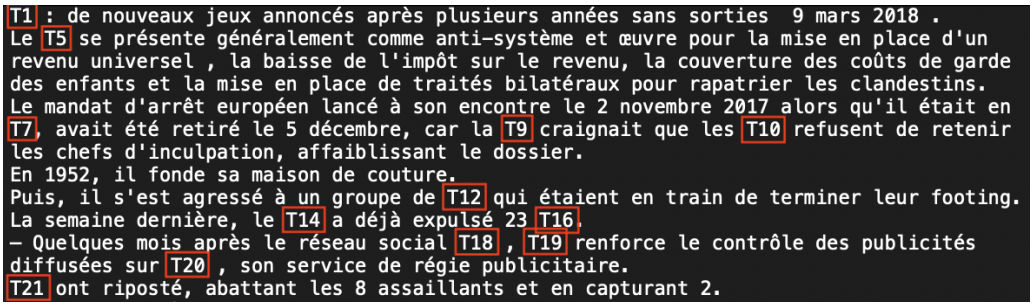

FIGURE 2.4 – Extrait du fichier *wikinews-2018-03_ condense.txt*.

Une fois ce premier travail sur les entités nommées fait, nous pouvions alors passer à l'échantillonnage de nos données.

2.2.3 Echantillonnage avec NLTK

Nous avons ensuite créé deux formes d'échantillons différents :
- des échantillons composés de texte et d'entités nommées comme dans les fichiers bruts ;
- des échantillons composés de texte avec à la place des entités nommées, leurs étiquettes (soit LOC, PER, ORG ou MISC).

Les échantillons étiquetés vont nous permettre de calculer les caractéristiques des entités nommées et les échantillons bruts, toutes les autres.

Utiliser l'expression régulière de NLTK au lieu de SemTagger à ce moment-là nous a permis de (re)tokeniser le texte comme : `['T1', 'où', 'les', 'T2', « l' », 'une', « l' », 'autre', « s' », 'attirent']`, et non comme : `['T', '1', 'où', 'les', 'T', '2', « l' », 'une', « l' », 'autre', « s' », 'attirent']`, le problème étant de ne pas pouvoir récupérer les entités nommées grâce à leur code.

Voici un extrait du résultat final (1. échantillons bruts et 2. échantillons étiquetés – passage du document *wikinews*) :

1. « *Puis , il s' est agressé à un groupe de* ***CRS*** *qui étaient en train de terminer leur footing . La semaine dernière , le* ***Royaume-Uni*** *a déjà expulsé 23* ***diplomates russes*** *. – Quelques mois après le réseau social* ***Facebook*** *,* ***Google*** *renforce le contrôle des publicités diffusées sur* ***AdWords*** *, son service de régie publicitaire .* ***Les forces de défense et de sécurité burkinabés*** *ont riposté , abattant les 8 assaillants* »

2. « *Puis , il s' est agressé à un groupe de* ***MISC*** *qui étaient en train de terminer leur footing . La semaine dernière , le* ***ORG*** *a déjà expulsé 23* ***MISC*** *. – Quelques mois après le réseau social* ***ORG*** *,* ***ORG*** *renforce le contrôle des publicités diffusées sur* ***MISC*** *, son service de régie publicitaire .* ***PER*** *ont riposté , abattant les 8 assaillants* »

Dans le tableau 2.1, nous pouvons voir le nombre d'échantillons créés pour chaque document de notre corpus, ainsi que leurs tailles exprimées en tokens. Cette préparation des données a permis d'obtenir des paires d'échantillons :
— de formes différentes (bruts et étiquetés) ;
— identiques en taille (même token de début et de fin) ;
— sans entités nommées segmentées dans deux échantillons différents.

Partant du postulat que nous avons besoin d'échantillons de tailles similaires pour notre expérience, nous avons décidé de ne conserver que les échantillons qui comptent plus de 190 tokens.

Nous avons donc obtenu 10 échantillons pour le genre de la prose (*pg6470* et *42131-0*), 10 pour celui de la poésie (*56708-0* et *pg6099*), 13 pour celui des informations (*GSD*, *Wikinews* et *APIL*), 5 pour le genre de l'encyclopédie (*aijwikiner*) et 5 pour celui de la parole transcrite (*spoken*).
Si l'on compare avec l'étude de Passonneau et al. (2014), ils avaient généré 965 échantillons de 500 tokens, soit environ 50 échantillons par genre.

Échantillons Fichiers	n°1	n°2	n°3	n°4	n°5	n°6
spoken	200	200	200	200	200	**40**
aijwikiner	200	200	197	200	200	**52**
GSD	200	199	200	200	**180**	/
wikinews	200	200	200	200	200	**41**
APIL	200	200	200	200	**112**	/
pg6470	200	200	200	200	190	**8**
42131-0	200	200	200	200	200	**59**
56708-0	200	200	200	200	200	**107**
pg6099	200	200	199	200	200	**46**
sequoia	200	200	200	200	**161**	/
FQB	200	198	200	200	200	**32**

TABLE 2.1 – Nombre d'échantillons par document et leurs tailles (en tokens).

Suite à cela, nous avons pu calculer nos caractéristiques.

2.3 Calculs des caractéristiques

2.3.1 Caractéristiques des entités nommées

Les premières caractéristiques (ou variables) que nous avons calculé sont celles des entités nommées, à partir des échantillons étiquetés :

- **feature_LOC** : la fréquence absolue des entités nommées de type LOC (*location*) par échantillon ;
- **feature_PER** : la fréquence absolue des entités nommées de type PER (*person*) par échantillon ;
- **feature_ORG** : la fréquence absolue des entités nommées de type ORG (*organization*) par échantillon ;
- **feature_MISC** : la fréquence absolue des entités nommées de type MISC (*miscellaneous*) par échantillon ;
- **feature_EN** : la fréquence absolue des entités nommées par échantillon, tous types confondus (LOC, PER, ORG, MISC).

	prose	parole	informations	encyclopedie	poesie	total
LOC	7	14	90	41	11	163
ORG	2	1	33	6	0	42
PER	46	5	20	22	22	115
MISC	22	8	49	17	17	113
TOTAL_EN	77	28	192	86	50	433
Nombre de tokens	1999	1000	2583	997	1999	8578

TABLE 2.2 – Nombre d'occurrences de chaque étiquette des entités nommées et nombre de tokens, par genre et tous genres confondus.

Au vu de la diversité d'entités nommées que la catégorie MISC regroupe :
— des professions ou fonctions, tels que : boucher, matelots, enchanteresse...
— des titres et noms en tous genres, tels que : Le Figaro, *New York Times*, *Hubble*, *Surface and Depth*...
— des évènements importants ou encore des traités, tels que : bataille d'Actium, accords sur le charbon et l'acier...

Par conséquent, nous avons décidé de ne pas prendre en compte cette catégorie d'entités nommées dans nos expériences, car nous ne serions en mesure d'interpréter précisément et de manière fiable l'impact qu'il aurait sur l'analyse en composantes principales.

2.3.2 Caractéristiques des parties du discours et des verbes au passé

Afin de mesurer l'importance de parties du discours (telles que les verbes, noms, déterminants, adjectifs...) et des verbes au passé comme descripteurs linguistiques dans la classification textuelle, nous avons procédé à une annotation automatique de notre corpus en deux temps :

1. pour l'annotation en parties du discours, nous avons testé plusieurs outils : 1) le modèle `POET` (*A French Extended Part-of-Speech Tagger*) de FLAIR[9], 2) CAMEMBERT[10] et 3) SPACY[11];

9. https://huggingface.co/qanastek/pos-french-camembert-flair
10. https://huggingface.co/gilf/french-camembert-postag-model
11. https://spacy.io/models/fr#fr_core_news_sm

2. pour l'annotation des verbes au passé, nous avons annoté notre corpus avec le *morphologizer* de SpaCy.

Pour l'étiquetage en partie du discours, après une brève comparaison des sorties produites par ces trois outils, nous avons choisi de nous baser sur l'annotation du modèle POET de Flair pour calculer nos caractéristiques.

En effet, CamemBERT :
— annotait souvent les espaces comme des signes de ponctuation (ex : `{'entity_group': 'PONCT', 'score': 0.91542035, 'word': '', 'start': 162, 'end': 163}`) ;
— pouvait scinder des segments annotables en deux et les annoter séparément, comme par exemple l'adjectif « intolérant » :
  — *{'entity_group' : 'ADJ', 'score' : 0.9674257, 'word' : '**intol**', 'start' : 141, 'end' : 147}*
  — et *{'entity_group' : 'NC', 'score' : 0.6571277, 'word' : '**érant**', 'start' : 147, 'end' : 152}*).

Quant à la qualité de l'étiquetage, CamemBERT pouvait annoter des signes de ponctuation par une autre partie du discours que PONCT, comme par exemple :
— un apostrophe par nom commun (NC) (*{'entity_group' : '**NC**', 'score' : 0.3736412, 'word' : « ' », 'start' : 56, 'end' : 57}*) ;
— un tiret par une conjonction de subordination *{'entity_group' : '**CS**', 'score' : 0.22250426, 'word' : '-', 'start' : 132, 'end' : 133}*).
... et bien d'autres, comme des chaînes de caractères telles que « (4me. », annotée comme un nom commun : {'entity_group' : '**NC**', 'score' : 0.77067095, 'word' : '(4me.', 'start' : 41, 'end' : 47}).

Le problème majeur de CamemBERT étant tout de même le découpage aléatoire de segments annotables, qui est relativement fréquent.

Concernant SpaCy, même si l'annotation en partie du discours semblait à première vue correcte, nous avons relevé que de nombreux verbes étaient annotés comme des noms (ex : « persécuté », « voudrions », « reprochons », « haïr », « prête »...) ou des adjectifs (ex : « brûla », « fais », « annonçaient »...), mais aussi des noms et des noms propres (ex : « vœu », « politique », « raisonneurs », « Luc »...).

Après cette étude rapide des sorties, nous avons donc choisi d'utiliser le modèle POET de Flair pour annoter notre corpus en partie du discours, et de SpaCy pour annoter nos verbes au passé. Voici un extrait d'une sortie de

FLAIR (document *wikinews*) :

```
Sentence[11]: « En outre, les USC Trojans et les UCLA Bruins figurent parmi les meilleures formations sportives universitaires américaines. » → [« En »/PREP, « outre »/ADV, « , »/PUNCT, « les »/DET, « USC »/PROPN, « Trojans »/XFAMIL, « et »/COCO, « les »/DET, « UCLA »/PROPN, « Bruins »/XFAMIL, « figurent »/VERB, « parmi »/PREP, « les »/ DET, « meilleures »/ADJFP, « formations »/NFP, « sportives »/ADJFP, « universitaires »/ADJFP, « américaines »/ADJFP, « . »/PUNCT]
```

... et d'une sortie de SPACY (document *pg6099*, poésie) :

```
senti  VERB  Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part
```

Après avoir nettoyé les sorties pour extraire les étiquettes générées, nous avons alors procédé à l'évaluation de l'annotation de FLAIR en parties du discours et de SPACY pour les verbes au passé.

2.3.3 Évaluation de l'annotation de FLAIR et SPACY

Hormis pour celles des entités nommées, le calcul de nos caractéristiques et par extension le résultat de notre analyse en composantes principales, se basent sur l'étiquetage en parties du discours de FLAIR, et de SPACY pour la caractéristique des verbes au passé. C'est pourquoi dans un souci de fiabilité, nous avons voulu brièvement vérifier manuellement les annotations produites par ces outils.

Pour évaluer FLAIR, nous avons sélectionné une centaine de tokens par document environ et observé les étiquettes. Le jeu d'étiquettes de ce modèle est le jeu d'étiquettes enrichi du corpus ANTILLES (*An Open French Linguistically Enriched Part-of-Speech Corpus*, cf. colonne 1 de la figure 2.5), basé sur le jeu d'étiquettes initial du corpus UD French-GSD. Nous avons regroupé nos étiquettes et évalué le modèle selon ce dernier.

UD French-GSD		ANTILLES	
Abbreviation	Description	Abbreviation	Description
ADP	Adposition	PREP	Preposition
		PART	Demonstrative particle
SCONJ	Subordinating conjunction	COSUB	Subordinating conjunction
CCONJ	Coordinating Conjunction	COCO	Coordinating Conjunction
ADV	Adverb	ADV	Adverb
PROPN	Proper noun	PROPN	Proper noun
		XFAMIL	Family name
NUM	Numerical Adjective	NUM	Numerical Adjective
		CHIF	Number
AUX	Auxiliary Verb	AUX	Auxiliary Verb
VERB	Verb	VERB	Verb
		VPP**XX** (x4)	FS/FP/MS/MP Past participle verb
		VPPRE	Present participle verb
DET	Determinant	DET	Determinant
		DET**XX** (x2)	FS/MS Determinant
ADJ	Adjective	ADJ	Adjective
		ADJ**XX** (x4)	FS/FP/MS/MP Adjective
		DINT**XX** (x2)	FS/MS Numerical adjectives
NOUN	Noun	NOUN	Noun
		N**XX** (x4)	FS/FP/MS/MP Noun
PRON	Pronoun	PRON	Pronoun
		PINT	FS Interrogative pronoun
		PDEM**XX** (x4)	FS/FP/MS/MP Demonstrative pronoun
		PIND**XX** (x4)	FS/FP/MS/MP Indefinite pronoun
		PPOBJ**XX** (x4)	FS/FP/MS/MP Pronoun complements of objects
		PPER1S	Personal pronoun - First person singular
		PPER2S	Personal pronoun - Second person singular
		PPER3**XX** (x4)	Personal Pronoun - Third Person FS/FP/MS/MP
		PREFS	Reflexive pronoun - First person of singular
		PREF	Reflexive pronoun - Third person of singular
		PREFP	Reflexive pronoun - First / Second Person of plurial
		PREL	Relative pronoun
		PREL**XX** (x4)	FS/FP/MS/MP Relative pronoun
INTJ	Interjection	INTJ	Interjection
SYM	Symbol	SYM	Symbol
PUNCT	Punctuation	YPFOR	Final point
		PUNCT	Punctuation
X	Other	MOTINC	Unknown word
		X	Typos & Other

FIGURE 2.5 – Jeu d'étiquettes enrichi du corpus ANTILLES (Labrak and Dufour, 2022), sur la base du jeu initial du corpus UD French-GSD.

Quant à SpaCy, les quatre type de temps annotés sont : *Past* (passé), *Imp* (imparfait), *Pres* (présent) et *Fut* (futur). Au vu du peu d'occurrences des annotations morphosyntaxiques *Tense=Past* (temps du passé, 274 tokens) et *Tense=Imp* (temps de l'imparfait, 84 tokens), nous avons relevé et évalué la précision de l'ensemble des tokens annotés comme tel (358 au total).

Modèle POET de Flair

	prose	parole	informations	encyclopedie	poesie	total
ADP	223	114	413	151	239	1140
SCONJ	44	18	9	3	23	97
CCONJ	42	42	64	26	80	254
ADV	134	118	79	30	98	459
PROPN	46	27	171	93	33	370
NUM	19	19	140	57	5	240
AUX	73	42	57	35	37	244
VERB	249	135	195	66	163	808
DET	250	104	367	139	323	1183
ADJ	219	118	317	137	312	1103
NOUN	309	154	584	192	438	1677
PRON	185	100	61	19	103	468
PPER1S	10	20	0	0	15	45
PPER2S	1	5	0	0	7	13
PPER3	50	16	20	7	14	107
INTJ	1	7	0	0	1	9
SYM	0	3	10	2	0	15
PUNCT	208	90	243	90	246	877
X	79	51	102	58	40	330
Nombre de tokens	1999	1000	2583	997	1999	8578

Table 2.3 – Nombre d'occurrences de chaque étiquette des parties du discours et nombre de tokens, par genre et tous genres confondus.

Dans la table 2.3, nous pouvons constater que la distribution des étiquettes est assez dispersée en général et entre les différentes catégories. Les interjections (9 occurrences), les pronoms personnels à la 2ème personne du singulier (13 occurrences) et les symboles (15 occurrences) sont peu nombreux, par rapport aux noms (1677 occurrences) ou les déterminants (1183

occurrences) par exemple. Nous pouvons penser qu'il y aura peut être un effet d'échantillonnage.

Globalement, nous pouvons constater que le modèle POET de FLAIR (table 2.4) est plus performant pour faire de l'étiquetage en parties du discours sur des types de textes encyclopédiques (94,64 % d'étiquettes correctes), informationnels (90,29 %), poétiques (90,14 %) et sur de la prose (88,56 %) que de la parole transcrite (78,43 %).

genre	document	échantillon (en tokens)	vrais positifs	précision / genre
parole	spoken	102	80	78,43
encyclopedie	aijwikiner	112	106	94,64
informations	gsd	95	86	
informations	wikinews	104	93	
informations	APIL	110	100	90,29
prose	pg6470	105	96	
prose	42131-0	96	82	88,56
poesie	56708-0	98	93	
poesie	pg6099	115	99	90,14

TABLE 2.4 – Résultat de l'évaluation de FLAIR, faite sur un échantillon d'une centaine de tokens environ par document.

Voici les erreurs les plus fréquentes que nous avons pu relever :

— FLAIR annote parfois les segments annotables séparés par des traits d'union comme un tout (ex :« post-punk »/NFS), parfois séparément (ex : « vert »/ADJFS, « - »/PUNCT, « de »/PREP, « - »/PUNCT, « gris »/ADJ) ;

— les occurrences du pronom personnel à la 1ère personne du pluriel « nous » sont tous annotés comme des pronoms réfléchis PREFS, même quand ils n'en sont pas ; en effet, ils pourraient être annotés PPER1P (à l'instar de PPER1S, mais cette étiquette n'existe pas dans le jeu)[12] ;

— beaucoup de signes de ponctuation (deux points « : », tirets « - », points « . », etc) sont étiquetés par d'autres parties du discours que PUNCT (ex : préposition (PREP), nom féminin singulier (NFS), déterminant (DET), etc) ;

12. Nous avons contacté les auteurs de l'article à ce sujet, mais notre question est restée sans réponse.

- la conjonction de subordination « comme » est souvent annoté comme une préposition (PREP) ;
- l'article partitif « du » et l'article indéfini « un » sont souvent annotés comme des adjectifs numéraux ordinaux (DINTMS) ;
- des erreurs causées par des fautes d'orthographe dans notre corpus, comme par exemple le document de la parole transcrite *spoken*, dans "voilà ou je vais", "ou" a été annoté comme une conjonction de coordination (CCONJ) au lieu d'un adverbe (ADV).

... et d'autres erreurs qui n'ont aucune incidence sur notre expérience, car il s'agit de confusions entre deux étiquettes du jeu enrichi d'ANTILLES, qui sont en fait réunies sous la même selon le jeu de l'UD French GSD, à savoir :

- des adjectifs numéraux (NUM, tels que : un, deux, trois...) sont annotés comme des chiffres (CHIF, tels que : 1, 2, 3...), par exemple dans le corpus : « cinquante »/CHIF ;
- idem pour les points finaux qui marquent la fin d'une phrase (YPFOR) et sont annotés comme des signes de ponctuation à part entière (PUNCT).

SpaCy

Globalement, nous pouvons constater que SpaCy (table 2.5) est plus performant sur les types de textes encyclopédiques (86,67 % d'étiquettes correctes), informationnels (77,42 %), sur de la prose (77,12 %), mais l'est beaucoup moins sur les types de textes oraux (58,14 %) et poétiques (37,29 %). Faute de temps, nous n'avons pas effectué une étude du rappel de l'outil, mais nous avons pu constater que certains verbes qui auraient dû être annotés au passé ont été annoté au présent (étiquette *Tense=Pres*, ex : « paya », « demanda » (document *pg6470*, prose), etc). Pour les verbes au passé de notre corpus, nous pouvons supposer un rappel bas et par conséquent, un grand silence.

genre	document	occurrences	vrais positifs	précision / genre
parole	spoken	43	25	58,14
encyclopedie	aijwikiner	45	39	86,67
informations	gsd	32	28	
informations	wikinews	33	31	
informations	APIL	28	13	77,42
prose	pg6470	61	49	
prose	42131-0	57	42	77,12
poesie	56708-0	29	17	
poesie	pg6099	30	15	37,29

TABLE 2.5 – Résultat de l'évaluation de SPACY, faite sur les 358 tokens annotés au temps du passé.

Voici les erreurs les plus fréquentes que nous avons pu relever :
— dans le document *spoken*, les interjections « euh » sont souvent annotés comme des verbes au passé ;
— dans les documents *APIL* (informations) et *56708-0* (poésie), beaucoup de noms dont la terminaison est « ée » sont annotés comme des verbes (par exemple, il y a confusion entre le nom « été » (la saison) et le participe passé de l'auxiliaire être (ex : « été » considéré comme un verbe du passé dans « Concours d'été »), le nom « randonnée », mais aussi des noms de villes (ex : BERULLE, ESTISSAC ...).

Ces erreurs, faute de temps, n'ont pas été corrigées manuellement par la suite, simplement relevées pour l'évaluation.

2.3.4 Choix des caractéristiques conservées

En procédant à l'évaluation manuelle et à l'exploration de ces annotations, nous avons décidé, à l'instar de la caractéristique des entités nommées de type MISC, de retirer de l'expérience les caractéristiques X et SYM :
— X, dont toutes les occurrences (23 dans tout le corpus) sont en fait des sous-étiquettes MOTINC de l'étiquette « X » qui correspondent à des mots inconnus, et sont essentiellement les mots anglais (ex : « *Practicing* », « *Philosophy* », « *Performing* », « *Live* »...), les consonnes euphoniques (ex : « t » dans « appelle-**t**-on ») et les disfluences de la parole transcrite (ex : « mh », « j » dans « **j**~**j**~**j**~ ») ;
— SYM, dont toutes les occurrences (18 dans tout le corpus) corres-

pondent à des symboles tels que le tilde (« ~ »), très présent notamment dans le genre de la parole, le pourcentage (« % »), les symboles monétaires (« € », « $ »), mais aussi à des abréviations et des symboles, comme « env » dans « **env**. deux fois plus élevés », « e » dans « 17**e** siècle », ou encore à des slashs « / » d'adresses URL (ex : dans « http ://villa-de-l-extra.skyrock.com/ »).

Nous ne les avons pas pris en compte car bien que ces étiquettes apparaissent peu de fois, ces caractéristiques pourraient avoir un impact trop fort sur l'ACP et biaiser nos résultats. D'autant qu'elles contiennent parfois des éléments bien différents pour que l'on puisse fournir une interprétation *a posteriori*.

Nous avons également ajouté la caractéristique de la longueur moyenne des tokens par échantillon.

Voici les caractéristiques sélectionnées pour l'ACP :

1. les entités nommées de lieux (LOC) ;
2. les entités nommées d'organisations (ORG) ;
3. les entités nommées de personnes (PER) ;
4. les entités nommées tous types confondus (TOTAL_EN) ;
5. la longueur moyenne des tokens par échantillon (LONGUEUR_MOTS) ;
6. les prépositions (PREP) ;
7. les conjonctions de surbordination (SCONJ) ;
8. les conjonctions de coordination (CCONJ) ;
9. les adverbes (ADV) ;
10. les noms propres (PROPN) ;
11. les nombres (NUM) ;
12. les auxiliaires (AUX) ;
13. les verbes (VERB) ;
14. les déterminants (DET) ;
15. les adjectifs (ADJ) ;
16. les noms (NOUN) ;
17. les pronoms (PRON) ;
18. les pronoms personnels à la 1ère personne du singulier (PPER1S) ;
19. les pronoms personnels à la 2ème personne du singulier (PPER2S) ;
20. les pronoms personnels à la 3ème personne (PPER3) ;

21. les interjections (INTJ) ;
22. les signes de ponctuation (PUNCT) ;
23. les verbes au passé (PAST_TENSE) ;

Ainsi, sur la base de 26 caractéristiques calculées, nous avons mené notre expérience sur 23 d'entre elles (sans MISC, X et SYM). Pour rappel, Passonneau et al. (2014) en avaient calculé au départ 37, pour n'en garder que 33.

Les caractéristiques que nous n'avons pas reprises de l'étude de Passonneau et al. (2014) sont celles des entités nommées sans date, des pronoms neutres *it*, des copules, des bases verbales, des participes présents et gérondifs, des participes passés, des verbes conjugués aux 1ère, 2ème personne et 3ème personnes du singulier, des superlatifs, des existentiels *there*, des verbes de persuasion, des verbes d'états, de la longueur des *chunks*[13] nominaux et des *chunks* verbaux, du nombre de tokens par phrase, du nombre de caractères par token, des points finaux, d'interrogations, d'exclamations et des virgules.

Certaines parce qu'elles peuvent être calculées à partir d'observables propres à la langue anglaise (notamment les pronoms neutres *it*, les existentiels *there* et la classe des verbes de persuasion (*suasive verbs*)) ; les autres faute de temps. Nous prévoyons d'en calculer une partie dans de futurs travaux.

Les échantillons et les observables ayant été définis, nous présentons dans la section suivante la technique statistique exploratoire que nous avons utilisé.

2.4 Analyse en Composantes Principales (ACP)

L'analyse en composantes principales est une méthode statistique exploratoire, qui permet de synthétiser un grand ensemble de données et d'en extraire les informations importantes. Elle réduit sa dimensionnalité en transformant des variables (ou caractéristiques) corrélées, en nouvelles variables décorrélées les unes des autres, et en les projetant dans un nouvel espace. Ces nouvelles variables sont appelées « composantes principales ». Il s'agit de combinaisons linéaires des variables d'origine qui expliquent le majeure partie de la variance

13. Les *chunks* en TAL sont des constituants continus et non-récursifs, qui définissent la structure syntaxique superficielle des phrases. (Tellier et al., 2012)

des données. La première composante principale explique le plus grande part de variance, la deuxième explique la deuxième plus grande part, et ainsi de suite.

Notre étude se base sur la distribution des descripteurs linguistiques, à savoir les caractéristiques que nous avons calculées précédemment, porteuses de certaines propriétés linguistiques de notre corpus.

2.4.1 Standardisation des données

Du fait de la grande hétérogénéité de nos variables, il était nécessaire de standardiser nos données, ceci dans le but de garantir qu'il n'y ait pas de déséquilibre dans la contribution des variables d'origine dû aux écarts entre les différentes valeurs.

En effet, l'ACP projette nos données originales dans des directions qui maximisent la variance. Par conséquent, si certaines variables ont une trop grande variance par rapport à d'autres, elle ne focalisera en quelque sorte que sur les grandes variances et les fera passer d'un faible impact (à l'échelle des autres) à la domination du premier composant principal.

Dans le tableau 2.6, nous pouvons constater que l'ordre de grandeur n'est pas le même pour toutes les caractéristiques.

TOTAL_EN	LONGUEUR_MOTS	PPER2S	VERB
0,025	4,555	0,0	0,14
0,03	4,635	0,0	0,1
0,03	4,835	0,0	0,115
0,02	4,44	0,0	0,135
0,02512	4,9497	0,0	0,12
0,04	4,475	0,005	0,125
0,02	4,89	0,0	0,155
0,015	4,58	0,005	0,1
0,025	4,8	0,005	0,145

TABLE 2.6 – Extrait de données des caractéristiques : entités nommées tous types confondus (TOTAL_EN), longueur des mots (LONGUEUR_MOTS), pronom personnel à la 2ème personne du singulier (PPER2S), verbes (VERB).

Après avoir chargé les données contenues dans un fichier csv, nous avons

ainsi standardisé nos données avec la fonction `StandardScaler` (librairie SCIKIT-LEARN) ; chaque variable est mise à l'échelle pour avoir une moyenne de zéro et un écart-type de un.

Puis, nous définissons le nombre de composantes avant d'appliquer la fonction `pca` (librairie SCIKIT-LEARN).

2.4.2 Choisir le nombre optimal de composantes

Il existe plusieurs moyens de déterminer quel est le nombre optimal de composantes pour l'ACP.

Le plus courant est de suivre la « règle de Kaiser-Guttman » qui est définie de la façon suivante : « *Le nombre des valeurs-propres supérieures à l'unité d'une matrice d'inter-corrélation*[14]*, est égal au nombre de facteurs à extraire* » (Ajar, 1982).

Ces valeurs-propres sont les *eigenvalues* et peuvent être obtenues avec le code `pca.explained_variance_`. Elles sont considérées comme une mesure de la variance expliquée de chaque composante principale.

Le principe est donc de définir autant de composantes qu'il y a de variables originales : le nombre de valeurs qui seront supérieures à 1 déterminera le nombre optimal de composantes à choisir.

Voici par exemple, les *eigenvalues* de nos 23 composantes :

```
[8.24003689 3.15024567 2.53492273 1.67165316 1.41529671
1.05434152 0.84525149 0.75280231 0.69398147 0.5986154 0.51233189
0.41240278 0.34888486 0.27946872 0.25968702 0.21869955
0.1608857 0.13160417 0.08728012 0.06572171 0.06028096 0.03732061
0.01590358]
```

Les 6 premières valeurs étant supérieures à 1, nous pouvons fixer le nombre de composantes principales à 6.

Mais cette règle pouvant être considérée comme « insuffisante », un autre moyen de valider ce nombre est de réaliser un *scree plot*.

Le *scree plot* (ou « éboulis ») introduit par Raymond B. Catell (1966), est un tracé linéaire des valeurs propres des facteurs ou des composantes principales d'une analyse. Elle permet d'afficher la proportion de variance expliquée pour chaque composante et de déterminer le nombre à retenir dans

14. Soit strictement supérieure à 1.

une analyse. L'idée est de détecter les « coudes » (les « cassures ») dans le graphique où les *eigenvalues* semblent dévier, avant de se stabiliser.

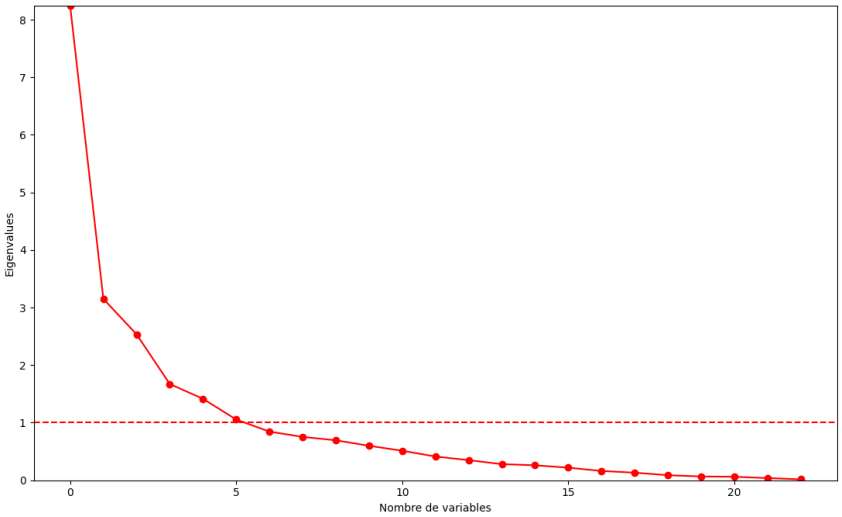

FIGURE 2.6 – *Scree plot* pour 23 variables.

Sur le *scree plot* de nos 23 variables en figure 2.6, nous pouvons observer deux coudes : en k = 2 et k = 4. Il semblerait que k = 6, pourtant à la limite du seuil de Kaiser (fixé à 1, *eigenvalue* à `1,05434152`), se trouve dans la zone de coude mais ne ressorte pas sur le graphique.

Pour préciser cette lecture, il nous semblait intéressant de compléter ce *scree plot* par un second graphique décrivant l'évolution de l'inertie expliquée par les axes, qui couplé au *scree plot* peut s'avérer décisif. Nous pouvons le retrouver en figure 2.7. Il s'agit de la courbe des valeurs propres cumulées, en pourcentage. Une analyse en k = 2 ne semble pas pertinente, car elles n'expliquent que 40 % de la variance. Le gain informationnel en intégrant les 4 axes suivants est d'environ 20 %. En effet, nous pouvons voir que k = 6 se trouve légèrement au-dessus de 80 % de variance expliquée ; au-delà, il est de plus en plus faible.

Ce deuxième graphique suggère qu'une analyse entre 4 et 6 composantes principales semble être le choix approprié pour nos données.

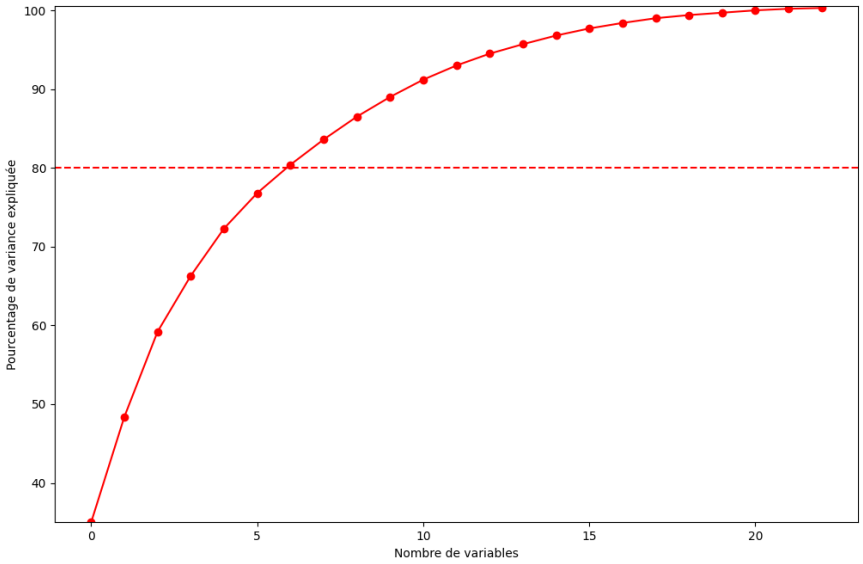

FIGURE 2.7 – Courbe des valeurs propres cumulées en pourcentage, pour 23 variables.

La section suivante présente les résultats de notre expérience.

Chapitre 3

Résultats

Dans ce chapitre consacré aux résultats, nous ferons la lecture de deux ACP différentes : une première basée sur les entités nommées annotées du corpus FENEC, et une deuxième en associant aux caractéristiques précédentes, celles des parties du discours et de la longueur des mots.

3.1 Exploitation des annotations existantes

La première ACP a été réalisée avec les caractéristiques disponibles sur le corpus FENEC, annoté notamment en entités nommées de lieux (LOC), de personnes (PER), d'organisations (ORG) et tous types confondus (TOTAL_EN).

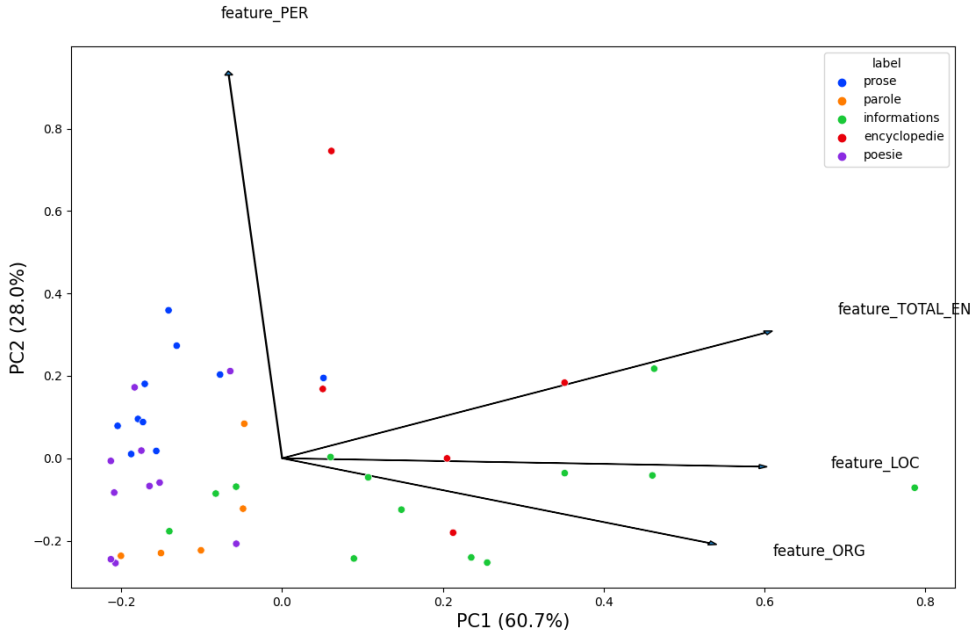

FIGURE 3.1 – 2 composantes principales, 4 caractéristiques (LOC, PER, ORG, TOTAL_EN).

Nous pouvons observer dans la figure 3.1 que la première composante se définit par les entités nommées de lieux (LOC), d'organisations (ORG) et les entités nommées tous types confondus (TOTAL_EN), tandis que la deuxième se définit par les entités nommées de personnes (PER).

Les échantillons des catégories de la prose et de la poésie se chevauchent, tandis que ceux des catégories de l'encyclopédie et des informations semblent éparpillés. Les échantillons de la parole se situent à peu près entre les deux groupes précédemment cités.

Bien que les entités nommées tous types confondus et plus précisément celles de lieux et d'organisations semblent plus caractéristiques des catégories de l'information et de l'encyclopédie, et les entités nommées de personnes des catégories de la prose et de la poésie, nous pouvons dire que dans le cadre de cette étude, les entités nommées à elles seules ne sont pas des descripteurs linguistiques suffisants pour discriminer les catégories textuelles.

3.2 Enrichissement des catégories

Dans l'étude de Passonneau et al. (2014), l'ACP a présenté quatre composantes principales. Tout comme eux, nous avons choisi de n'afficher que les caractéristiques qui ont des *loadings* supérieurs à 0,2 (valeur absolue). Les *loadings* correspondent aux poids de l'association entre les variables d'origine et les composantes principales ; autrement dit, il s'agit des caractéristiques qui définissent le plus une composante principale.

3.2.1 Première et deuxième composantes principales

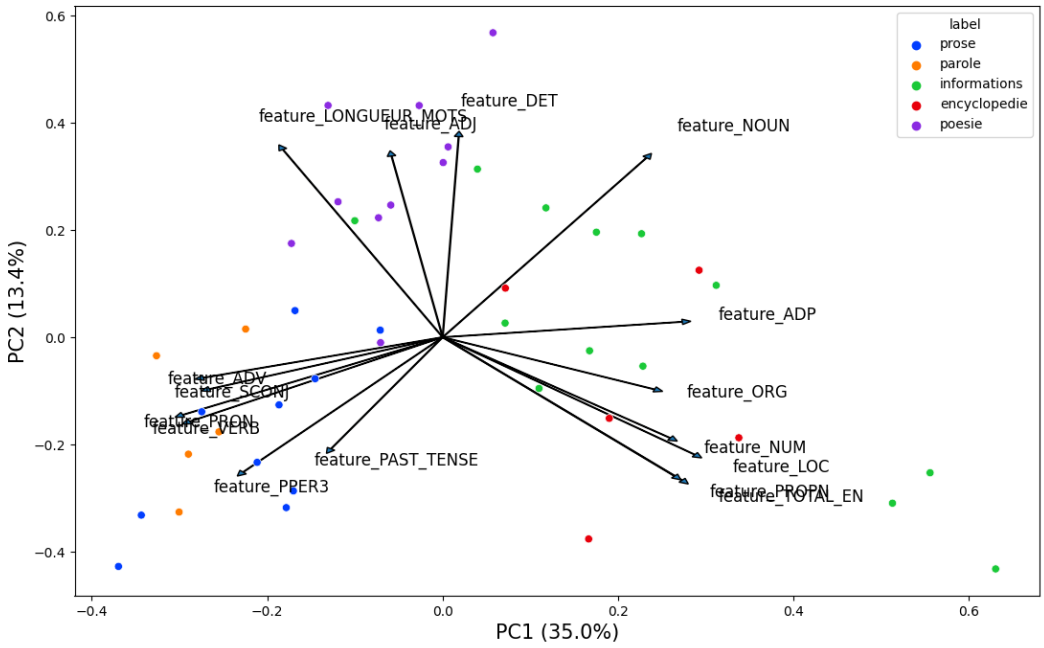

FIGURE 3.2 – Première et deuxième composantes principales, 23 caractéristiques.

Comme nous pouvons le voir sur la figure 3.2, notre première composante principale se définit majoritairement :

— à gauche, par les caractéristiques des adverbes (ADV), des conjonctions de subordination (SCONJ), des pronoms (PRON), des pronoms person-

nels à la 3ème personne (PPER3), des verbes (VERB), des verbes au passé (PAST_TENSE) et dans la diagonale, de la longueur des mots (LONGUEUR_MOTS) ; de ce coté sont regroupés les échantillons des catégories de la prose et de la parole ;

— à droite, par les caractéristiques des prépositions (ADP), des nombres (NUM), des noms propres (PROPN), des entités nommées tous types confondus (TOTAL_EN), notamment les entités nommées de lieux (LOC), d'organisations (ORG) et dans la diagonale, des **noms** (NOUN) ; de ce côté se trouvent certains échantillons des catégories de l'information et de l'encyclopédie.

Les caractéristiques qui contribuent le plus à notre deuxième composante principale sont à nouveau celles des noms (NOUN) et de la longueur des mots (LONGUEUR_MOTS) (qui contribuent aux deux composantes), mais aussi des entités nommées de personnes (PER), des adjectifs (ADJ) et des déterminants (DET). De ce côté sont regroupés les échantillons de la poésie, avec quelques échantillons d'informations.

Dans la diagonale inverse se trouve les caractéristiques des noms propres (PROPN), des entités nommées tous types confondus (TOTAL_EN), là où se trouvent des échantillons des catégories de l'information et de l'encyclopédie.

Dans notre PCA, deux dimensions semblent se dessiner. D'une part, les textes de la poésie sont plus caractérisés par des mots plus longs et des noms détaillés, précisés par des adjectifs ; ceux de la prose et de la parole, par des conjonctions de subordination, des verbes en général, des verbes au passé, et à la voix passive. D'autre part, nous pouvons penser que les textes d'informations et d'encyclopédies présentent plus d'éléments tels que des zones géographiques, lieux d'intérêt, dates historiques, numéros de téléphones, sommes d'argent, etc.

La quatrième composante de Passonneau et al. (2014) correspondait à la cinquième de Biber (1988), qu'il identifiait comme la dimension « abstrait *VS* non-abstrait ». De leur côté, la composante se définissait surtout par les caractéristiques des virgules, des **prépositions**, de la longueur des phrases (en tokens), des **participes passés** et des **verbes de base**. Dans l'étude de Biber (1988), il s'agissait des conjonctions, ce qui d'après Passonneau et al. (2014), pourrait être corrélé avec des phrases plus longues, des participes passés, et des voix passives sans agent. La dispersion globale de leurs échantillons était plus uniforme dans les deux dimensions, avec des centres séparés pour chacune des catégories mais encore une fois, sans séparation nette.

De plus, la deuxième composante de Passonneau et al. (2014) était définie

presque entièrement par les **entités nommées** et les **noms communs**, et ne correspondait à aucune des composantes de Biber (1988) car il n'avait pas de caractéristiques d'entités nommées. Sur le nuage de points, ils ont observé qu'il existait des régions distinctes, mais avec de nombreux chevauchements, et que deux catégories se démarquaient. Chaque genre (pour rappel, trouvés par regroupement hiérarchique : la parole, la narration, la production écrite et discursif) présentait une plus grande dispersion le long de la deuxième composante. La fréquence des entités nommées variait selon les textes de ces catégories, à l'exception des deux catégories singletons – guides de voyage (similaire à notre document APIL du corpus) et documents techniques – qui elles se trouvaient à l'extrême du genre discursif, mais toutes deux opposés. Ces dernières incluaient systématiquement plus d'entités nommées.

Nos première et deuxième composantes présentent donc des similitudes avec ces dernières. En effet, nous avons retrouvé des catégories de productions écrites, détaillées d'un côté (poésie, prose), soit la dimension abstraite ; et des catégories explicatives de l'autre (informations, encyclopédie), soit la dimension non-abstraite. Dans le cadre de cette étude, ces composantes ne sont malgré tout pas suffisantes pour discriminer la parole de la prose.

3.2.2 Troisième et quatrième composantes principales

Sur la figure 3.3, nous pouvons voir que les caractéristiques qui ont le plus de poids pour notre troisième composante principale sont :

- à gauche, ceux des **pronoms personnels à la 3ème personne** (PPER3), de la longueur des mots (LONGUEUR_MOTS), des déterminants (DET), des **verbes au passé** (PAST_TENSE), des auxiliaires (AUX), des prépositions (ADP) et des entités nommées d'organisations (ORG) ; là se trouvent des échantillons de l'information, de la prose et quelques uns de l'encyclopédie et de la poésie ;
- à droite, ceux de la ponctuation (PUNCT) (là où sont majoritairement regroupés des échantillons de la poésie), ceux des conjonctions de coordination (CCONJ), des pronoms personnels à la 1ère (PPER1S) et à la 2ème personne du singulier (PPER2S), et enfin ceux des interjections (INTJ) ; c'est notamment dans ces directions que nous retrouvons des échantillons de la parole.

Des échantillons de l'information sont éparpillés dans les deux directions, dans l'axe de cette troisième composante.

Notre quatrième composante semble se définir par les caractéristiques des entités nommées de personnes (PER), mais aussi des interjections (INTJ) et de la **ponctuation** (PUNCT), ces deux dernières ayant du poids à la fois dans

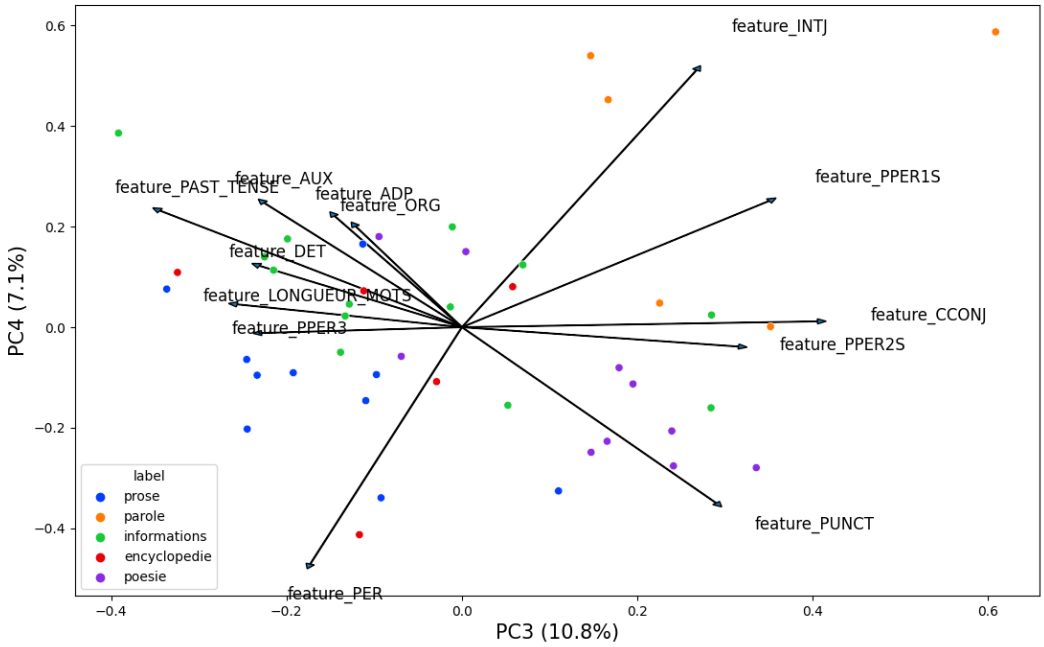

FIGURE 3.3 – Troisième et quatrième composantes principales, 23 caractéristiques.

notre troisième et quatrième composante.

Dans notre PCA, deux dimensions semblent à nouveau se dessiner. D'une part, les textes de parole transcrite qui se caractérisent par des interactions en face à face (pronoms personnels aux 1ère et 2ème personne du singulier), avec des interjections et des conjonctions de coordination. D'autre part, les textes en prose présentent plus d'entités nommées de personnes, de pronoms à la 3ème personne, et de ponctuation. Dans les textes d'informations, nous pouvons trouver de la narration d'évènements passés, avec des mots plus longs et plus de noms d'organismes.

La première composante de Passonneau et al. (2014) correspondait à la première de Biber (1993), qu'il interprétait comme la dimension « impliqué *VS* informationnel » (soit des données non planifiées interactives, principalement orales contre des écrits (parfois dense) qui informent sur un sujet donné). Les caractéristiques qui contribuaient le plus directement à leur composante étaient les **verbes au passé**, les copules, les **pronoms personnels**,

les adverbes, le **nombre de caractères par mot**, les noms, et pour celle de Biber (1988), il s'agissait des **pronoms personnels**, des adverbes, des noms et de la **longueur des mots**. Les documents d'interactions orales dans MASC se situaient côté "impliqué" de cette dimension, et les textes explicatifs côté "informationnel".

Aussi, la troisième composante de Passonneau et al. (2014) correspondait à la dimension narrative de la deuxième composante de Biber (1988) qu'il interprétait comme la dimension « narratif VS non-narratif ». Les caractéristiques qui avaient le plus de poids pour leur composante étaient majoritairement des formes verbales (notamment les **verbes au passé**, mais aussi les verbes au présent à la 3ème personne), des **pronoms à la 3ème personne** et des **entités nommées de personnes**. Dans l'étude de Biber (1988), il s'agissait plus des **verbes au passé** et des **pronoms à la 3ème personne**, et moins des verbes au présent.

Nos troisième et quatrième composantes présentent donc des similitudes avec ces dernières. En effet, nous avons retrouvé des textes d'interactions orales d'un côté (parole), soit la dimension impliquée ; et textes explicatifs de l'autre (informations, prose), soit la dimension informationnelle (ou narrative). En revanche, dans le cadre de cette étude, ces composantes ne sont pas suffisantes pour discriminer plus nettement chaque catégorie textuelle, notamment celle de l'encyclopédie.

3.2.3 Cinquième et sixième composantes principales

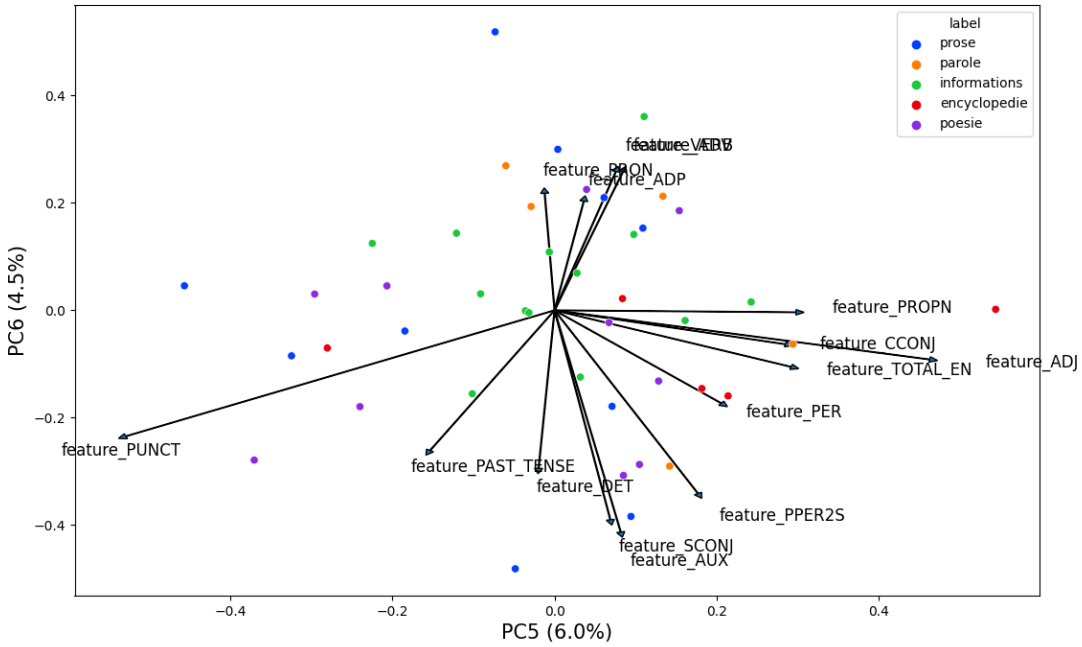

FIGURE 3.4 – Cinquième et sixième composantes principales, 23 caractéristiques.

Sur la figure 3.4, notre cinquième composante semble se définir par des signes de ponctuation (PUNCT) à gauche, et par des noms propres (PROPN), des entités nommées de personnes (PER), des entités nommées tous types confondus (TOTAL_EN), des conjonctions de coordination (CCONJ) et des adjectifs (ADJ).

Notre sixième composante quant à elle, semble se définir par des pronoms (PRON), des verbes (VERB), des prépositions (ADP), et des adverbes (ADV) d'un côté, et par des pronoms personnels à la 2ème personne du singulier (PPER2S), des auxiliaires (AUX), des verbes au passé (PAST_TENSE), des déterminants (DET), des conjonctions de subordination (SCONJ).

La lecture de nos deux dernières composantes est difficile. En effet, celles-ci expliquent très peu la variance (6,0 % pour PC5 et 4,5 % pour PC6). En effet, tous les échantillons sont éparpillés, avec une concentration d'informations au centre et d'encyclopédie à droite. La prose se trouve surtout aux extrémités (haut, bas, gauche), et la parole (haut, bas, droite). Nous ne

retrouvons aucune similitude avec les six composantes de Biber (1988) et les quatre de Passonneau et al. (2014).

En conclusion, notre expérience a permis, en utilisant un corpus de petite taille et des caractéristiques relativement facile à calculer, de retrouver certaines des dimensions proposées par Biber (1988) et Passonneau et al. (2014) : les dimensions « abstrait VS non-abstrait », « impliqué VS informationnel » et « narratif VS non-narratif ».

Conclusion

En conclusion, ce travail nous permet d'avoir une vision plus riche de l'influence des catégories textuelles traditionnelles (genres/domaines/modalités/registres) et de faire certaines observations.

Tout d'abord, cette expérience nous a permis de nous confronter à la difficulté de tokeniser proprement et uniformément un ensemble de documents hétérogènes. En effet, nous avons vu que les outils peuvent avoir des comportements imprévisibles et différents sur un même phénomène linguistique, et au sein d'un même texte.

Ensuite, bien qu'il reste une notion de continuum à approfondir, les résultats de la première ACP nous permettent d'observer que dans le cadre de cette expérience, les entités nommées de lieux, de personnes et d'organisations ne sont pas des caractéristiques suffisantes pour catégoriser à elles seules les textes.

En ajoutant les caractéristiques d'autres descripteurs linguistiques telles que les parties du discours, la longueur des mots et les verbes au passé, nous avons pu retrouver certaines similitudes avec les dimensions textuelles que Biber (1988) et Passonneau et al. (2014) avaient fait émerger en anglais britannique et américain. Mais pour la première fois, sur un jeu de données multi-catégories français.

Aussi, la troisième composante de Passonneau et al. (2014) correspondait à la dimension narrative de la deuxième composante de Biber (1988) qu'il interprétait comme la dimension « narratif VS non-narratif ».

En effet, nos première et deuxième composantes présentent des similitudes avec les cinquième de Biber (1988) et quatrième de Passonneau et al. (2014) (abstrait VS non-abstrait). Les catégories de productions écrites, détaillées d'un côté (poésie, prose) semblent appartenir à la dimension abstraite ; et des catégories explicatives de l'autre (informations, encyclopédie), à la dimension non-abstraite.

Nos troisième et quatrième composantes présentent des similitudes avec les premières dimensions de Biber (1988) et Passonneau et al. (2014) (impliqué VS informationnel), et leurs deuxième (Biber, 1988) et troisième (Passon-

neau et al., 2014) dimensions (narratif VS non-narratif). Les textes d'interactions orales d'un côté (parole) semblent appartenir à la dimension impliquée ; et les textes explicatifs de l'autre (informations, prose), à la dimension informationnelle (ou narrative, avec les conjonctions, participes passés, et voix passive).

Malgré ces similitudes, la taille déjà peu conséquente de notre corpus (seulement deux ou trois documents par catégorie textuelle de 1 000 tokens environ), appauvri de deux documents pour des questions d'interprétabilité des données, produit un effet d'échantillonnage fort ; de ce fait, la taille de nos échantillons et la faible variance de nos dernières composantes ne nous permettent pas de faire une analyse plus décisive de celles-ci. A l'avenir, nous prévoyons d'augmenter la taille du corpus FENEC en ajoutant des nouveaux documents annotés en entités nommées, et de faire varier la taille des échantillons pour voir si celle-ci peut impacter les résultats de l'expérience.

Nous pouvons également discuter de l'impact qu'a la qualité de l'annotation des systèmes que nous avons utilisés pour annoter notre corpus en parties du discours et en traits morphosyntaxiques, sur nos résultats. Au total, nous avons testé trois outils : CamemBERT et le modèle POET de Flair pour l'annotation en parties du discours, et SpaCy pour les deux types d'annotation. L'évaluation de ces derniers nous a permis de constater sur la tâche d'étiquetage en parties du discours, ce qu'avaient observé Millour et al. (2022) sur la tâche de REN pour le français : un différentiel de performances important entre les catégories textuelles, avec des faiblesses majeures sur des textes de parole transcrite et de poésie.

Bien que ces annotations soient dans l'ensemble correctes, n'ayant pu les corriger manuellement faute de temps, nous reconnaissons que les variables calculées à partir de ces dernières en ont sûrement été impactées, et par conséquent, nos composantes principales. Nous projetons à l'avenir de collaborer au cours d'une thèse, avec le « Consortium CORpus, Langues et Interactions » (CORLI) qui a lancé le projet « Annotation » en 2023. Les annotations sont réalisées par des annotateurs humains (*a minima* une licence en linguistique), qui s'appuient sur un guide d'annotation et utilisent l'outil INCEPTION[1]. Les annotations se font sur plusieurs couches et sont de natures très diverses : catégories morpho-syntaxiques, annotations sémantiques ou discursives ; dans le cas de corpus oraux ou multi-modaux, informations sur la prosodie, gestes, etc. Le cadre expérimental mis en place dans notre étude est plutôt modulable et permet facilement d'enrichir l'expérience avec

1. https://inception-project.github.io/

d'autres caractéristiques, d'autres types de textes, et de donner lieu à de nombreuses extensions. Rattachées à l'initiative du projet « Annotation », nous pourrons ainsi travailler à partir d'un ensemble d'échantillons hétérogène riche en annotations validées manuellement, afin de trouver de nouveaux indices quant à la caractérisation textuelle.

En effet, dans nos analyses en composantes principales, nous avons toujours du mal à distinguer les textes de type encyclopédie et informations, or nous pouvons observer un réel différentiel de performances sur les tâches d'étiquetage morpho-syntaxique[2] et de REN (Millour et al., 2022). Une meilleure description de ces catégories nous permettrait de faire une évaluation plus interprétable.

Nous pouvons par exemple faire l'hypothèse que les performances des outils sont moindres sur le type de texte de la parole transcrite spontanée, en raison de structures syntaxiques bouleversées par des amorces, des tics de langage, des répétitions, ou encore des auto-interruptions (par exemple, « ouais donc euh tu vois c'est ça euh que j'ai vu »). Nous pourrions aussi faire l'hypothèse que ce même outil aura des performances équivalentes sur des textes issus du réseau social *X* (anciennement *Twitter*). En effet, ces micro-textes dans des séries d'interactions soumis à des règles textuelles (nombre limité de caractères), sont vecteurs d'une nouvelle forme d'écriture depuis plusieurs années : syntaxiquement (*#hashtags* en milieux et fins de phrases...), lexicalement (*emojis* à la place des mots, lieu d'apparition de nouveaux mots d'argots et expressions...), etc.

A l'avenir, Millour et al. (2022) prévoient d'agrandir le corpus FENEC avec deux documents de parole transcrite de 1 000 tokens environ (issu du corpus Rhapsodie Lacheret, Anne et al. (2014)), un document en prose de 1 000 tokens (« L'Homme qui plantait des arbres » de Giono) et un document journalistique de 1 000 tokens (du journal « L'Est Républicain »), tous annotés en entités nommées. Une des extensions envisageables de cette étude est de s'inspirer des travaux de Fu et al. (2020) et d'approfondir la notion des caractéristiques des entités nommées. Car au-delà de leurs types, il s'agit sûrement leurs propriétés (longueur en caractères et en tokens, fréquence, ambiguïté, persistance...) qui mettent en difficulté les outils.

2. **Modèle POET de Flair** : encyclopédie (précision : 94,64 %), informations (précision : 90,29 %) / **SpaCy** : encyclopédie (précision : 86,67 %), informations (précision : 77,42 %)

Annexes

Extraits pour chaque document du corpus

aijwikiner (genre : encyclopédie)

Deux temps structurent l' œuvre, les cinquante premiers livres sont dédiés aux 723 premières années de Rome, de sa fondation à la bataille d' Actium en -31 et les trente derniers aux 250 années impériales. L' Inde a trois périodes de vacances nationales. On retrouve l' influence de Ballard dans la musique de groupes de post-punk comme Joy Division, The Normal et John Foxx. Jiang Qing, la femme de Mao Zedong, et la bande des Quatre agite le mouvement contre les chaînes culturelles du passé : de nombreuses œuvres anciennes, livres, sculptures, bâtiments, etc. L' explication d' Arrhenius est que lors de sa dissolution, le sel se dissocie en particules chargées (que Michael Faraday avait nommé « ions » quelques années avant). Plus récemment, la série Fringe se déroule également à Boston. Les populaires séries télévisées québécoises Lance et compte (1986 –) ont également le hockey sur glace comme thème central.

APIL (genre : informations)

ERVY-LE-CHATEL – 2ème jour du Marché du Livre à la salle des fêtes de 10 h à 18 h. Rencontres avec les auteurs, lectures entre 12 h et 14 h. Organisée par « Ervy-le-Châtel, village du livre et des arts » 03 25 76 88 78. 23 juillet : LIGNIERES – Reconnaître 5 sauvageonnes comestibles. Savoir les utiliser et les cuisiner. Organisé par Les Ombelles 03 25 43 92 26 ou www.lesombelles.com Octobre 2011 15 septembre : Sortie au départ d'AIX-EN-OTHE – Sortie au Musée Colette et Château de Guédelon. Organisé par le Club des Anciens et de amis d'Aix 06 75 15 06 14. 10 et 11 décembre : BOUILLY – Marché de Noël au foyer familial, organisé par le Comité des fêtes. ERVY-LE-CHATEL – Récital de Piano 4 mains avec Matthieu Normand et Nicolas Collinet à la Halle Circulaire à 20 h 30. Organisé

par Villa de l'Extra 03 25 43 22 10 ou 06 79 07 79 25 ou http ://villa-de-l-extra.skyrock.com/

FQB (genre : multi-sources)

Où fut ouvert le premier magasin J. C. Penney ? Combien de gallons d'eau y a-t-il dans un pied cube ? À quelle date est né Dwight D. Eisenhower ? Quelle est la population du Maryland ? Comment appelle-t-on un groupe d'antilopes ? Que symbolisent les « flèches circulaires » ? Qui a été le troisième président des États-Unis ? Quelle planète est la plus éloignée du soleil ? Dans quelle province se trouve Montréal ? Quand le téléphone fut-il inventé ? Woodrow Wilson a assumé la présidence de quelle université ? Dans quelle nation se trouve la Kaaba ? Comment l'obtenir ? Qui fait la météo dans l'émission télévisée « Good Morning America » ? Sous quel surnom le musicien Ernesto Antonio Puente Junior était-il mieux connu ? Que sont les enzymes ? Combien de gens écoutent la télévision en réseau ? En quelle année Mozart est-il né ? Combien d'îles forment l'Indonésie ?

GSD (genre : multi-sources)

Outre ces îles, la ville est composée de nombreux quartiers tels que La Capte, Giens, L'Almanarre, L'Ayguade, Le Pyanet, Costebelle, Les Salins-d'Hyères ou Les Borrels. D'un côté El Niño est responsable d'une importante sécheresse et de l'autre une vague de froid traverse le pays. Shusterman poursuit cette réflexion dans plusieurs essais subséquents tels Practicing Philosophy (1997), Performing Live (2000) et Surface and Depth (2002). Le ‘ ndranghetiste, du mot grec Andragatos, qui signifie homme valeureux et courageux, a des origines lointaines. Cet antisémitisme ne constitue pas l'ensemble de la réflexion d'un penseur qui est l'un des plus grands analystes du XXème siècle en France. Et puis, ma chaîne est tombée. Puis elle rejoint Hollywood où, de 1930 à 1942, elle participe à 175 films américains au sein de la Metro-Goldwyn-Mayer, principalement comme conceptrice de robes, entre autres aux côtés d'Adrian. La Centrale féline Belge (CFB) est un registre d'élevage belge.

sequoia (genre : multi-sources)

Expression parue dans la revue n° 33 « Verdun la vie » et envoyée à tous les Verdunois, ne correspond pas à la réalité, ce que soulignent M. Christian Langlois et M. Patrice Lanini, délégués syndicaux CGT des Rapides de la Meuse. Les multiples auditions conduites par le juge Armand Riberolles

ont essentiellement porté sur le rôle du promoteur Jean-Claude Méry, dont les révélations avaient relancé l'enquête. Plusieurs hommes auraient, pour le compte de Francis Poullain, fait plusieurs aller et retour entre la France et l'Afrique pour transporter des fonds. François Mitterrand était alors ministre des Colonies (à partir de 1950) et René Bousquet directeur de la Banque d'Indochine. - Le 17 mai 2004, l'ancien ministre socialiste du Budget Michel Charasse a reconnu avoir signé des commissions « légales » pour les frégates de Taïwan. 5 x 1 flacon (conditionnement unitaire) 2 flacons Quelques toiles, de l'acrylique, de l'huile, de l'encre de Chine, mais aussi de la sculpture et bien d'autres techniques. Blanche Maupas, décédée en 1962, avait aussi écrit un livre, paru en 1933, Le Fusillé, dont la réédition en 1994 comporte des illustrations de Tardi.

spoken (genre : parole)

Expression parue dans la revue n° 33 « Verdun la vie » et envoyée à tous les Verdunois, ne correspond pas à la réalité, ce que soulignent M. Christian Langlois et M. Patrice Lanini, délégués syndicaux CGT des Rapides de la Meuse. Les multiples auditions conduites par le juge Armand Riberolles ont essentiellement porté sur le rôle du promoteur Jean-Claude Méry, dont les révélations avaient relancé l'enquête. Plusieurs hommes auraient, pour le compte de Francis Poullain, fait plusieurs aller et retour entre la France et l'Afrique pour transporter des fonds. François Mitterrand était alors ministre des Colonies (à partir de 1950) et René Bousquet directeur de la Banque d'Indochine. - Le 17 mai 2004, l'ancien ministre socialiste du Budget Michel Charasse a reconnu avoir signé des commissions « légales » pour les frégates de Taïwan. 5 x 1 flacon (conditionnement unitaire) 2 flacons Quelques toiles, de l'acrylique, de l'huile, de l'encre de Chine, mais aussi de la sculpture et bien d'autres techniques. Blanche Maupas, décédée en 1962, avait aussi écrit un livre, paru en 1933, Le Fusillé, dont la réédition en 1994 comporte des illustrations de Tardi. Nous devons considérer les personnes qui votent pour ces partis et comprendre pourquoi cette situation voit le jour.

pg6099 (genre : poésie)

Lesbos où les Phrynés l'une l'autre s'attirent, Où jamais un soupir ne resta sans écho, A l'égal de Paphos les étoiles t'admirent, Et Vénus à bon droit peut jalouser Sapho ! –Lesbos où les Phrynés l'une l'autre s'attirent. –Cependant tu vas gueusant Quelque vieux débris gisant Au seuil de quelque Véfour De carrefour ; Je te frapperai sans colère Et sans haine,–comme un boucher ! Comme Moïse le rocher, –Et je ferai de ta paupière, Ils trottent,

tout pareils à des marionnettes ; Se traînent, comme font les animaux blessés, Ou dansent, sans vouloir danser, pauvres sonnettes Où se pend un Démon sans pitié ! Tout cassés Son teint est pâle et chaud ; la brune enchanteresse A dans le col des airs noblement maniérés ; Grande et svelte en marchant comme une chasseresse, Son sourire est tranquille et ses yeux assurés. Tu vas lorgnant en dessous Des bijoux de vingt-neuf sous Dont je ne puis, oh ! pardon ! Te faire don ; –Plus belle que Vénus se dressant sur le monde Et versant les trésors de sa sérénité Et le rayonnement de sa jeunesse blonde Sur le vieil Océan de sa fille enchanté ; Plus belle que Vénus se dressant sur le monde ! Et pourtant aimez-moi, tendre coeur ! soyez mère Même pour un ingrat, même pour un méchant ; Amante ou soeur, soyez la douceur éphémère D'un glorieux automne ou d'un soleil couchant.

pg6470 (genre : prose)

Les mauvaises pensées me dérangent trop. – Toi, mon petit, je t'en.... et, tiens ! Ce fut Alexandre qui paya. grand poëte que vous êtes ! demanda la vieille, toute frétillante, enchantée d'apprendre que les deux femmes s'étaient disputées. Elle était tout près de lui. Les paroles de Lisa retentissaient, comme s'il eût déjà entendu les fortes bottes des gendarmes, à la porte de la chambre. J'y arriverai, à l'histoire du monsieur... Je te raconte l'histoire tout entière. Elle voulut deux tranches de galantine ; elle aimait ça. Mais elle dit que non, qu'il fallait savoir où ils en étaient auparavant. Tous trois s'en allaient, traînant les talons sur les trottoirs, tenant la largeur, forçant les gens à descendre. Alors, mademoiselle Saget lui répondait que, dame ! Si, d'ailleurs, on ne l'avait pas arrête, au 2 décembre, ce n'était pas sa faute. Tenez, je ne vous aime plus. Cette brutalité jeta Florent hors de lui. – Merci, gronda la portière, cinquante francs, pour l'avoir dorloté avec de la tisane et du bouillon ! C'était barbare et superbe, quelque chose comme un ventre aperçu dans une gloire, mais avec une cruauté de touche, un emportement de raillerie tels, que la foule s'attroupa devant la vitrine, inquiétée par cet étalage qui flambait si rudement... Quand ma tante Lisa revint de la cuisine, elle eut peur, s'imaginant que j'avais mis le feu aux graisses de la boutique. Là, dans l'air renfermé, dans le demi-jour des quelques becs de gaz, il retrouvait la fraîcheur de l'eau pure. – Monsieur Gavard est tout an fond, dit le jeune homme, qui marchait toujours.

wikinews (genre : informations)

Valve : de nouveaux jeux annoncés après plusieurs années sans sorties 9 mars 2018 . Le M5S se présente généralement comme anti-système et œuvre

pour la mise en place d'un revenu universel , la baisse de l'impôt sur le revenu, la couverture des coûts de garde des enfants et la mise en place de traités bilatéraux pour rapatrier les clandestins. Le mandat d'arrêt européen lancé à son encontre le 2 novembre 2017 alors qu'il était en Belgique, avait été retiré le 5 décembre, car la justice espagnole craignait que les juges belges refusent de retenir les chefs d'inculpation, affaiblissant le dossier. En 1952, il fonde sa maison de couture. Puis, il s'est agressé à un groupe de CRS qui étaient en train de terminer leur footing. La semaine dernière, le Royaume-Uni a déjà expulsé 23 diplomates russes. – Quelques mois après le réseau social Facebook , Google renforce le contrôle des publicités diffusées sur AdWords , son service de régie publicitaire. Les forces de défense et de sécurité burkinabés ont riposté, abattant les 8 assaillants et en capturant 2. Valve recommence à publier des jeux. Plus de cinquante wikimédien.ne.s provenant de plusieurs pays africains, Algérie , Égypte , Cameroun , Nigéria etc. – Le Glacier Perito Moreno classé au Patrimoine mondial de l'UNESCO [1] est sur le point de s'effondrer. Il s'agit d'un homme marocain de 26 ans qui aurait déclaré agir au nom de Daesh . La rentrée atmosphérique de Tiangong-1 intéresse particulièrement l'agence européenne, mais aussi les autres agences spatiales.

42131-0 (genre : prose)

Et quand nous leur reprochons d'avoir persécuté, voudrions-nous être persécuteurs ? Le même Isaïe marche tout nud, pour marquer que le Roi d'Assyrie emmenera d'Egypte et d'Ethiopie une foule de captifs qui n'auront pas de quoi couvrir leur nudité. Vertu vaut mieux que science. On dirait qu'on a fait vœu de haïr ses freres ; car nous avons assez de religion pour haïr et persécuter, nous n'en avons pas assez pour aimer et pour secourir. L'Ecriture nous apprend donc que non-seulement Dieu tolérait tous les autres Peuples, mais qu'il en avait un soin paternel : et nous osons être intolérants ! Elle arriva à Paris prête d'expirer. Epictete dans les fers, Marc-Antonin sur le Trône, disent la même chose en cent endroits. [20] On ne révoque point en doute la mort de St. Ignace ; mais qu'on lise la Relation de son martyre, un homme de bon sens ne sentira-t-il pas quelques doutes s'élever dans son esprit ? Qu'on ne fasse aucune violence aux Juifs, (4me. Pierre-le-Grand a favorisé tous les Cultes dans son vaste Empire : le Commerce et l'Agriculture y ont gagné, et le Corps politique n'en a jamais souffert. St. Luc en compte quarante-une ; et ces générations sont absolument différentes. Il est dit que dans la guerre qu'il fit aux Madianites, [25]Moïse ordonna de tuer tous les enfants mâles et toutes les meres, et de partager le butin. (Le Cardinal le Camus, Instruction pastorale de 1688.) En attendant, je ne puis que remercier la Providence de ce qu'elle permet que les gens de son espece soient toujours de mauvais

raisonneurs. Le motif de l'arrêt était aussi inconcevable que tout le reste.

56708-0 (genre : poésie)

Et le Splendide-Hôtel fut bâti dans le chaos de glaces et de nuit du pôle. Dames qui tournoient sur les terrasses voisines de la mer : enfantes et géantes, superbes noires dans la mousse vert-de-gris, bijoux debout sur le sol gras des bosquets et des jardinets dégelés,–jeunes mères et grandes sœurs aux regards pleins de pèlerinages, sultanes, princesses de démarche et de costume tyranniques, petites étrangères et personnes doucement malheureuses. La musique, virement des gouffres et chocs des glaçons aux astres. Nous faisions un tour dans la banlieue. L'aube d'or et la soirée frissonnante trouvent notre brick au large en face de cette villa et de ses dépendances qui forment un promontoire aussi étendu que l'Épire et le Péloponèse, ou que la grande île du Japon, ou que l'Arabie ! Les airs et les formes mourant...–Un chœur, pour calmer l'impuissance et l'absence ! Ce poison va rester dans toutes nos veines même quand, la fanfare tournant, nous serons rendu à l'ancienne inharmonie. Le peuple ne murmura pas. Il ne faut même plus songer à cela. Les desperadoes languissent après l'orage, l'ivresse et les blessures. Ils se pâmaient l'un contre l'autre. Quand nous sommes très forts,–qui recule ? Depuis lors, la Lune entendit les chacals piaulant par les déserts de thym,–et les églogues en sabots grognant dans le verger. Le moment de l'étuve, des mers enlevées, des embrasements souterrains, de la planète emportée, et des exterminations conséquentes, certitudes si peu malignement indiquées dans la Bible et par les Normes et qu'il sera donné à l'être sérieux de surveiller.–Cependant ce ne sera point un effet de légende ! Le Prince était le Génie. Nymphes d'Horace coiffées au Premier Empire.–Rondes sibériennes, Chinoises de Boucher.

Bibliographie

Djavid Ajar. Le problème de la détermination du nombre de facteurs en analyse factorielle. Revue des sciences de l'éducation, 8(1) :45–62, 1982. doi : https://doi.org/10.7202/900356ar.

Douglas Biber. Variation across Speech and Writing. Cambridge University Press, 1988. doi : 10.1017/CBO9780511621024.

Douglas Biber. Representativeness in corpus design. Literary and linguistic computing, 8(4) :243–257, 1993.

Éric Bidaud and Hakima Megherbi. De l'oral à l'écrit. La lettre de l enfance et de l adolescence, 61, 01 2005. doi : 10.3917/lett.061.24.

Marie Candito and Djamé Seddah. Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical (the sequoia corpus : Syntactic annotation and use for a parser lexical domain adaptation method) [in French]. In Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN, pages 321–334, Grenoble, France, June 2012. ATALA/AFCP. URL https://aclanthology.org/F12-2024.

Jinlan Fu, Pengfei Liu, and Graham Neubig. Interpretable multi-dataset evaluation for named entity recognition. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6058–6069, Online, November 2020. Association for Computational Linguistics. doi : 10.18653/v1/2020.emnlp-main.489. URL https://aclanthology.org/2020.emnlp-main.489.

Yanis Labrak and Richard Dufour. ANTILLES : An Open French Linguistically Enriched Part-of-Speech Corpus. In 25th International Conference on Text, Speech and Dialogue (TSD), Brno, Czech Republic, September 2022. URL https://hal.science/hal-03696042.

Lacheret, Anne, Kahane, Sylvain, Beliao, Julie, Dister, Anne, Gerdes, Kim, Goldman, Jean-Philippe, Obin, Nicolas, Pietrandrea, Paola, and Tchobanov, Atanas. Rhapsodie : un treebank annoté pour l´étude de l´interface syntaxe-prosodie en français parlé. SHS Web of Conferences, 8 :2675–2689, 2014. doi : 10.1051/shsconf/20140801305. URL https://doi.org/10.1051/shsconf/20140801305.

Pengfei Liu, Jinlan Fu, Yang Xiao, Weizhe Yuan, Shuaichen Chang, Junqi Dai, Yixin Liu, Zihuiwen Ye, and Graham Neubig. ExplainaBoard : An explainable leaderboard for NLP. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing : System Demonstrations, pages 280–289, Online, August 2021. Association for Computational Linguistics. doi : 10.18653/v1/2021.acl-demo.34. URL https://aclanthology.org/2021.acl-demo.34.

Denise Malrieu and François Rastier. Genres et variations morphosyntaxiques. Traitements automatiques du langage, 16, 2001.

Alice Millour, Yoann Dupont, Alexane Jouglar, and Karën Fort. FENEC : un corpus équilibré pour l'évaluation des entités nommées en français (FENEC : a balanced sample corpus for French named entity recognition). In Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale, pages 82–94, Avignon, France, 6 2022. ATALA. URL https://aclanthology.org/2022.jeptalnrecital-taln.8.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. Learning multilingual named entity recognition from wikipedia. Artificial Intelligence, 194 :151–175, 2013. ISSN 0004-3702. doi : https://doi.org/10.1016/j.artint.2012.03.006. URL https://www.sciencedirect.com/science/article/pii/S0004370212000276. Artificial Intelligence, Wikipedia and Semi-Structured Resources.

Rebecca J. Passonneau, Nancy Ide, Songqiao Su, and Jesse Stuart. Biber redux : Reconsidering dimensions of variation in American English. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics : Technical Papers, pages 565–576, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics. URL https://aclanthology.org/C14-1054.

Céline Poudat, Guillaume Cleuziou, and Viviane Clavier. Catégorisation de textes en domaines et genres : complémentarité des indexations lexi-

cale et morphosyntaxique. Document numérique - Revue des sciences et technologies de l'information. Série Document numérique, 9 :61–76, 2006. doi : 10.3166/dn.9.1.61-76. URL https://hal.science/hal-00084803.

T. B. W. Reid. Linguistics, structuralism, and philology. Archivum Linguisticum, 8 :28–37, 1956.

Djamé Seddah and Marie Candito. Hard time parsing questions : Building a QuestionBank for French. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 2366–2370, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1375.

Isabelle Tellier, Denys Duchier, Iris Eshkol, Arnaud Courmet, and Mathieu Martinet. Apprentissage automatique d'un chunker pour le français. In Georges Antoniadis, Hervé Blanchon, and Gilles Sérasset, editors, TALN2012, volume 2 of Actes de la conférence conjointe JEP-TALN-RECITAL 2012, pages 431–438, Grenoble, France, June 2012. URL https://hal.science/hal-01174591.