Projet GénéAlgoPsy : Analyse exploratoire des emplois associés aux douleurs physique et mentale dans un corpus anglophone

Lou Gellad

Stage de fin de troisième année de licence

Institut Villebon-Georges Charpak - Université Paris-Saclay

14 avril-31 juillet 2024 au laboratoire LATTICE ENS CNRS PSL

Résumé

Suite à une analyse longitudinale sémantique et syntaxiques des emplois liés aux douleurs physique et mentale sur plusieurs siècles, faite par le laboratoire LATTICE en partant du corpus Gallica de la BnF, j'ai été chargée de constitué un corpus comparable à partir de COHA dans le but de vérifier si l'on pouvait trouver les mêmes tendances dans les mêmes langues. De nombreux problèmes sont intervenus avec l'utilisation de COHA et les méthodes d'analyse employées pour Gallica ne sont majoritairement pas exploitable sur COHA. Des voies d'amélioration sont envisagées et il sera possible d'agglomérer plusieurs corpus anglophones pour avoir un nombre conséquent d'occurrences.

Mots clés : douleur, douleur mentale, douleur sociale, dépression, suicide, analyse culturelle, résultats, santé mentale, mélancolie, humanités numériques, humanités médicales

Abréviation	Thème	Explicitation
GénéAlgoPsy	Projet	Généalogie Culturelle de la Douleur Psychique : Expressions, Usages, Représentations
LATTICE	Laboratoire	Langues, Textes, Traitements informatiques et Cognition
CRESS	Centre de Recherche	Centre de Recherche en Epidémiologie et Statistiques
METHODS	Equipe	Méthodes d'évaluation thérapeutique des maladies chroniques
ENS	Centre de Recherche	Ecole Normale Supérieure
CNRS	Centre de Recherche	Centre National de la Recherche Scientifique
BnF	Bibliothèque	Bibliothèque Nationale de France
COHA	Corpus	Corpus of Historical American English
WLP	Nom de fichier	
textID	Colonne	Identifiant de l'ouvrage
BNC	Traitement de données	British National Corpus
text	Nom de fichier	Type de table de COHA (voir 2. COHA)
spacy	Nom de package	
UMR	Terme administratif	Unité Mixte de Recherche
SGML	Langage balisé	Standard Generalized Markup Language
XML	Langage balisé	Extensible Markup Language
DTD	Type de fichiers	document type definition
DocBook	Type de balisage	-
RAINBOW	Type de balisage	-
TEI	Type de balisage	Text Encoding Initiative
CES	Type de balisage	Corpus Encoding Standard
zip	Mode de compression	move at high speed
gzip	Mode de compression	-
FIPAT	Fédération de recherche	The Federative International Programme for Anatomical Terminology
TNA	Table	Terminologia NeuroAnatomica
TA2	Table	Terminologia Anatomica 2
CNRC-NRC	Centre de recherche	The National Research Council of Canada Industrial Research Assistance Program

Table des matières 3

Remerciements

Le stage au sein de ce laboratoire restera un excellent souvenir grâce à la cohésion de l'équipe et du dévouement de mes superviseuses Frédérique et Laure, ainsi que de Qi, Yoann et Mathieu, sans qui, je ne serai jamais parvenue à de tels résultats.

Etant pourtant partie avec une appréhension sur ma capacité à tenir dans la durée, j'ai réussi à me faire une place au sein de ce laboratoire chaleureux et dynamique, ambiance permise grâce à tous ces membres où associés qui y gravitent.

Je voulais également souligner la patience de Frédérique, notamment quand j'emploi « base de données » à la place de « table », mais dans bien d'autres cas encore. Laure a également su s'adapter à ma demande de faire plus de linguistique et a su se mettre à mon niveau, m'expliquer et me faire confiance et je l'en suis très reconnaissante.

1. Introduction

Contexte de l'étude - Présentation du laboratoire

Ce rapport concerne la période de mon stage de fin de licence effectué en présentiel, du 15 avril au 18 juin 2024, au sein du laboratoire LATTICE, situé à l'Ecole Normale Supérieure de Montrouge dans les Haut-de-Seine. Le LATTICE (Langues, Textes, Traitements informatiques et Cognition) est un laboratoire d'enseignement et de recherche sous la triple tutelle du CNRS, de l'Ecole Normale supérieure et de l'Université Paris Sorbonne-Nouvelle. Il abrite des chercheurs du CNRS, des enseignants-chercheurs de l'Université Paris Sorbonne-Nouvelle et de d'autres universités, des ingénieurs d'études, des post-doctorants, des doctorants, d'autres membres associés ainsi que des stagiaires (niveau licence et master).

L'unité de recherche a été créée en janvier 2000 et se consacre à la linguistique et au traitement automatique des langues. A l'interface de ces deux domaines se développent des projets de recherche qui répondent aux axes de recherches du laboratoire que sont le lexique, le discours, le changement linguistique et l'évolution des langues, la modélisation, le traitement automatique et les applications.

Durant les deux premiers mois de ce stage, j'ai travaillé sous la tutelle de Laure Sarda, chargée de recherche en linguistique, et de Frédérique Mélanie-Becquet, ingénieure d'études. Frédérique Mélanie-Becquet, ma principale interlocutrice, est spécialisée en production, traitement et analyse de données, et membre du Comité du réseau Mate-SHS (Réseau professionnel de partage de méthodes d'analyses de terrains et d'enquête en SHS). Laure Sarda est, elle, une linguiste spécialisée dans la description linguistique de l'espace. J'ai eu également d'autres interlocuteurs internes ou externes au laboratoire, comme présentés en annexe.

Contexte de l'étude - Présentation du projet et enjeux

Le projet GénéAlgoPsy correspond à l'intitulé « Généalogie Culturelle de la Douleur Psychique : Expressions, Usages, Représentations ». Il est porté par Astrid Chevance, psychiatre et chargée de recherche en Epidémiologie et Statistiques dans l'équipe METHODS du CRESS (Centre of Research in Epidemiology and StatisticS) UMR 11553. Ces travaux s'inscrivent dans un programme de recherche interdisciplinaire sur quatre ans qui vise à décrire et définir le concept de la douleur, et plus particulièrement celui de la « douleur mentale » dans une perspective thérapeutique.

Dans cette optique, la cheffe de projet puise des ressources dans plusieurs domaines de recherches : la médecine et biologie, l'anthropologie sociale, la philosophie analytique, et enfin la linguistique de corpus et traitement de données.

Contexte de l'étude - Son objectif

La douleur mentale est un symptôme transdiagnostique pouvant prédire des tendances suicidaires. [1] C'est également le symptôme considéré comme le plus dur à vivre pour les patients. [2] Cependant, la recherche clinique autour de sa définition et de sa mesure reste superficielle. Elle est rarement évaluée et traitée lors du suivi des patients, contrairement à la douleur physique. [3] Pourtant la douleur mentale n'est pas une notion nouvelle puisqu'elle est déjà évoquée dans des textes aliénistes européens du XIXe siècle (Guillain, Séglas, Maudsley, Krafft-Ebing) sous la forme de caractéristique centrale de la folie ou de mélancolie.

Ainsi, afin de mieux comprendre l'expression des douleurs physiques et mentales au fil des siècles, l'étude dans laquelle mon stage prend place s'inscrit dans une démarche historique et linguistique, ayant pour objectif de retracer l'évolution de l'expression, des usages et des représentations de la notion de douleur mentale liés aux mots « douleur » en français et « pain » en anglais depuis l'époque moderne (1650) jusqu'à aujourd'hui (2021).

A la suite d'un appel à projet réalisé en juin 2022, GénéAlgoPsy a été l'un des quatre projets sélectionnés par la BnF DataLab pour un accueil et accompagnement scientifique, technique et financier sur une durée d'un an. La BnF est la Bibliothèque Nationale de France, et le DataLab est un « service à destination des chercheurs qui souhaitent travailler sur les collections numériques de la BnF ».

Contexte de l'étude - Travaux préliminaires

Avant mon arrivée et avant toute étude de corpus, l'équipe avait déjà entamé des premiers travaux de réflexion sur la construction d'un corpus, et la catégorisation du type douleur entre physique et psychique.

Afin d'évaluer la faisabilité du projet de recherche, une étude pilote a été mise en place. Un corpus de 40 000 occurrences des lemmes « douleur » ou « douloureux » sur une période entre 1650 et 1980 a été construit à partir d'une base de données de textes en ligne (Frantext). Chaque occurrence est mise sous format de concordancier : 600 caractères avant, le mot pivot et 600 caractères après, afin de conserver les informations sémantiques et syntaxiques du contexte des extraits d'ouvrages.

Depuis ce corpus, 1005 occurrences ont été extraites afin d'être annotées manuellement selon la catégorie de la douleur exprimée. Nous distinguons la douleur physique de la douleur psychique, les contextes médicaux de non médicaux. Ces données annotées ont ensuite servi de corpus d'entraînement afin d'entraîner un modèle de classification automatique classant les ouvrages selon ces deux axes.

Contexte de l'étude - Gallica

Le corpus francophone est issu de Gallica, la bibliothèque numérique de la BnF. Il est composé de toutes les occurrences du nom « douleur » et de l'adjectif « douloureux », avec un contexte de 600 caractères avant et 600 caractères après ainsi que les métadonnées correspondantes (auteur, titre du document, date de publication et domaine de la source (ex : médecine, littérature, philosophie…)), afin de coller au corpus d'entrainement.

Suite à la constitution de ce corpus, il a paru nécessaire de savoir si les analyses faites en français pouvaient être généralisable à l'anglais. Mais la constitution d'un corpus anglophone n'est pas une mince affaire.

Les défis actuels de la création de grands corpus

Avec la prolifération des textes numériques et la numérisation des œuvres imprimées par des institutions publiques comme la BNF et des entreprises technologiques comme Google, la création de vastes corpus textuels ne semble plus être un défi. Pourtant, le traitement efficace de ces immenses quantités de données brutes et de leurs annotations reste un travail fastidieux et complexe.

En effet, la création de corpus reste une tâche technique puisqu'il faut être en capacité de jongler entre données brutes, métadonnées et annotations de façon à ce qu'aucune donnée ne soit abusivement supprimée et que les trois types de données soient reliées correctement par un identifiant unique. Il faut également être en capacité de trouver ou créer des outils d'annotation efficients qui soient adaptés au type de données (langue, registre de langue, époque…) dont la puissance soit compatible avec les outils du laboratoire.

Tout d'abord, chaque caractère du texte correspond à une retranscription en octet suivant certaines conventions comme ASCII, UNICODE ou ISO-Latin-1. L'encodage définit donc la méthode d'ouverture d'un fichier pour le programme qui doit le faire. Certains programmes n'ouvrent que certains encodages et il nécessaire de choisir un seul encodage d'arrivée pour le corpus, pourtant il n'y a pas de norme internationale. Il faut donc essayer de choisir l'encodage qui sera le plus compatible avec les programmes utilisés dans le ou les pays cible(s) du corpus. Par exemple, ISO est censée être la norme internationale mais elle est très peu utilisée au japon et dans les pays arabes. L'encodage conseillé dans la majorité des cas est l'UTF-8 car il convient à un large public et cette norme est plus constante dans le temps.

De plus, l'exploitation de données textuelles en vue de leur valorisation scientifique et pécuniaire doit se conformer aux problématiques éthiques et législatives. Ainsi, la constitution de corpus impose de naviguer entre les restrictions imposées par les droits de propriété intellectuelle et de protection des données personnelles.

Une fois les textes rassemblés, s'ajoute le défi de l'annotation. En effet, dans la plupart des cas, l'annotation linguistique automatique ne peut se substituer à l'annotation manuelle. L'étiquetage des données fait par des modèles informatiques spécialisés sont parfois trop inexactes et certaines tâches, notamment sémantiques nécessitent une interprétation humaine difficilement informatisable. Or, il est indispensable de garantir la précision et la cohérence des annotations pour que les analyses qui en découlent soient fiables.

Ensuite, la constitution de corpus n'est pas une finalité : le formatage des données a avant tout l'objectif de simplifier leur traitement ultérieur. Se pose alors la question de l'interrogation efficace de très grands corpus. Le format doit être compatible avec différents langages informatiques, nécessiter des processeurs et de la mémoire vive raisonnables tout en permettant de façon intuitive l'analyse des données.

Deux alternatives majoritaires de forme de stockage de données existent pour répondre à ces problématiques : la base de données et le langage balisé. Les bases de données relationnelles ou non peuvent être stockées dans des formats textes divers (json, xml). Elles sont le plus souvent exportables sous format tabulaire, compatible avec de nombreux logiciels couramment utilisés comme Microsoft Excel ou Lotus. Elle simplifie les calculs statistiques et rend accessible la visualisation des données via les outils intégrés aux tableurs. Cependant, il peut s'avérer difficile de passer d'une base de données à un autre format de données, le choix de ce format est donc une décision majeure au sein d'un projet de recherche.

La deuxième option (présentée ci-dessous) utilise des langages comme le SGML et le XML et est de plus en plus employée, notamment aux Etats-Unis, car elle a le gros avantage d'être compatible avec Internet. Dans ce type de format, chaque région du texte est encadrée par

des balises d'ouvertures et de fermetures qui contiennent un ou plusieurs qualificatif(s) de la donnée.

```
<quotation>To be or not to be.</quotation>
<quotation author='WS'>. . .</quotation>
```

Figure 1. Exemple de langage balisé

Il existe plusieurs jeux de Définitions de Type de Document (DTD) pour réglementer l'usage des balises qui sont généralement spécialisées dans un type d'application. DOCBOOK et RAINBOW sont utilisées pour créer des documents électroniques alors que TEI et CES sont davantage adaptés) au marquage de textes existants afin de produire des corpus généralement utiles à des fins de recherche. Jusqu'à il y a peu de temps, la technique des DTD obligeait le constructeur du corpus à créer un logiciel propre permettant d'exploiter son balisage. Cependant, hormis dans le cas de balisages « faits mains », il existe de plus en plus de logiciels libres compatibles avec les DTD standards comme WIN32 et UN*X, si bien que cette limitation est en train de disparaitre.

Le format CoNLL-U

Pour le corpus francophone comme pour le corpus anglophone, après plusieurs changements de trajectoire, l'équipe a choisi le format CoNLL-U comme format standard, principalement pour l'affinité des chercheurs avec celui-ci et parce qu'il nous permettait de rassembler au même endroit contexte, syntaxe et découpage en mots, ce qui nous simplifie grandement les analyses.

Il impose un encodage en UTF-8 et doit contenir un identifiant correspondant au rang du premier caractère du mot dans la phrase, le lemme (c'est-à-dire le masculin singulier ou l'infinitif), l'étiquetage morpho-syntaxique de chaque mot avec l'Universal part-of-speech tag (c'est-à-dire l'étiquetage grammatical du mot), la tête (c'est-à-dire le mot qui est qualifié par l'indication donnée par notre mot : l'adjectif a pour tête un nom par exemple), la relation de dépendance (COD, COI, attribut, sujet …) et d'autres informations complémentaires.

Il est bien-sûr possible de moduler les informations en fonction de l'orientation du corpus. [4]

Enfin, la question la plus épineuse liée à la distribution est la compression de données. Il faut déjà se poser la question de s'il on souhaite compresser ou non les données en fonction de la taille et de l'utilisation que l'on souhaite faire du corpus. Si l'on décide de passer par une compression, la question de la portabilité, c'est-à-dire de l'accessibilité des données, se pose comme elle se posait avant pour l'encodage. En effet, les bibliothèques libres de compression et décompression de fichiers comme zip et gzip ne sont pas universelles et certaines plateformes ne tolèrent que certains modes de compression. Le choix du mode de compression a également un impact sur la vitesse de décompression et donc sur l'accessibilité des données. [5][6]

Le choix de l'orientation du corpus

Contrairement à ce que l'on pourrait penser, un corpus n'est pas généraliste. Si l'on propose des ouvrages répartis sur une longue période, le but est d'étudier l'évolution dans le temps de quelque chose. Il peut s'agir également de corpus centrés sur un mot pivot, dans ce cas-là, il ne sera pas possible de ne pas travailler sur ce pivot sans être biaisé. C'est également le cas pour les corpus de traduction : il y a souvent une langue pivot par laquelle passe tous les alignements (c'est-à-dire la correspondance entre un groupe de mot dans une langue et

dans une autre). Ainsi, apporter des informations, des axes de recherche à un corpus, c'est aussi lui donner un certain positionnement. [7]

Pour Généalgopsy, il fallait donc que le corpus anglophone est une approche longitudinale pour pouvoir faire des études sur les structures autour de la notion de douleur à travers les siècles. Il fallait également avoir le nom des auteurs, les dates et les genres, parce qu'ils étaient des paramètres d'étude du corpus francophone. Et enfin, il fallait transformer ce corpus en un nouveau, qui soit plus spécialisé, afin de correspondre à la problématique de recherche d'Astrid Chevance.

2. Matériel

2.1 COHA

Le **C**orpus **o**f **H**istorical **A**merican English (COHA), développé par la Brigham Young University, est une sélection de textes historiques anglais provenant de journaux, de magazines, de fictions, de textes académiques et de scénarios de films publiés entre 1810 et 2009. [8][9]

Ce corpus a été choisi parce qu'il est très utilisé en linguistique pour tirer des exemples contenant un pivot via le site internet. Des personnes de l'équipe étaient donc convaincues qu'il s'agirait du meilleur corpus pour notre tâche.

Il est disponible gratuitement en ligne [10], cependant, le nombre de requêtes quotidiennes est extrêmement limité. Un extrait du corpus est téléchargeable en ligne [11] et enfin une version payante de COHA à 375$ par an [8] permet d'y accéder de façon illimité via le téléchargement des données. Cette dernière version est composée de deux types principaux de table et d'un fichier de données que nous allons décrire présentement. [13]

Les lexiques (ou wlp)

lexicon table

wordID	word	lemma	PoS
71186	swab	swab	nn1
77653	swag	swag	nn1
36155	swagger	swagger	nn1

Figure 2 Les lexiques

Les tables wlp (comme ci-dessus) contiennent les informations sur chaque mot de chaque texte par genre et décennie. La première colonne correspond à l'identifiant de l'ouvrage (ou textID), le mot, le lemme et la classe grammaticale selon le British National Corpus (ou BNC). [14]

Le lemme est la forme du mot sans ses possibles marques de flexion. On peut également l'appeler « forme canonique » ou « forme dictionnairique ». Par exemple, « fleur » est le lemme de « fleurs », « envoyer » celui de « enverrons » et « passionnant » celui de « passionnante ». [15]

Les textes linéaires (ou text)

Les fichiers text contiennent tous les ouvrages d'une décennie pour un genre donné, mis bout à bout. Chaque ouvrage est délimité au début par son textID précédé de deux arrobases et à la fin par un retour à la ligne. Ainsi, une ligne dans un bloc-notes (sans retour à la ligne automatique) équivaut à un ouvrage. Le genre et la décennie ne sont pas mentionnés dans le fichier, seul l'identifiant du texte l'est, et ces informations ne sont disponibles que dans le titre du document.

Le fichier « sources »

COHA

textID	Year	genre	sourceTitle	textTitle
728282	1837	FIC	Source_A	SampleTitle_N
728283	1872	FIC	Source_B	SampleTitle_0
728284	1904	NF	Source_C	SampleTitle_P
728285	1938	MAG	Source_D	SampleTitle_Q
728286	1959	NEWS	Source_E	SampleTitle_R
728287	1987	MAG	Source_F	SampleTitle_S

Figure 3 COHA sources

Ce dernier fichier unique (comme ci-dessus) regroupe l'ensemble des informations pour chaque textID.

La première colonne correspond à l'identifiant du texte, la deuxième au nombre de mots dans l'ouvrage, la troisième au genre, la quatrième à la date, la cinquième au titre et la dernière contient des informations complémentaires.

2.2 Les limites de COHA

Même si COHA offre de nombreuses possibilités, ce corpus n'est pas sans de nombreuses limitations.

Non normalisation des tables : séparateurs, colonnes manquantes et encodage

Les séparateurs

Dans un tableur, les séparateurs permettent de délimités le passage d'une colonne à une autre au sein d'une même ligne d'un fichier. Cependant, une diversité de séparateurs coexiste dans le fichier source, comme « / », «
 » (le retour à la ligne) ou « \t » (la tabulation) et des lignes sont parfois sautées dans la base de données. De plus, les séparateurs peuvent aussi être utilisés au sein d'une même cellule ce qui rend difficile voire impossible une importation de la base de donnée (voir ci-dessous). J'ai par conséquent décidé de supprimer toutes les balises «
 », les sauts de lignes et les « / » et de choisir comme séparateur unique la tabulation, le point-virgule étant déjà utilisé dans les titres des ouvrages.

	textID	nbOfWords	genre	yearPub	title	extra
Description colonne	Identifiant ouvrage	Nb. mots ouvrage COHA	Fiction, Journal…	Année de publication	Titre ouvrage	Informations complément.
Ligne dans COHA	« Or, the Steam Man of the Prairies »	Ellis, Edward Sylvester, 1840-1916	\n [retour à la ligne]	[vide]	[vide]	[vide]

Figure 4 Exemple de problème de délimitation de ligne dû au mauvais usage de séparateurs

Les colonnes manquantes

Les pratiques ne sont pas unifiées mais dans la plupart du temps, si une colonne n'est pas remplie, elle est supprimée pour la ligne. En d'autres termes, la valeur de la troisième colonne d'une ligne peut être placée dans la deuxième colonne si la deuxième colonne est vide. La quatrième colonne sera dans la troisième et ainsi de suite.

Cela a deux problèmes majoritaires. Le premier est que nous ne pouvons pas ouvrir de façon rapide la table dans Python puisque les librairies ne tolèrent que très peu les tables non carrées, c'est-à-dire avec un nombre irrégulier de colonnes par ligne.

Le deuxième problème est que nous n'avons aucune indication sur la colonne manquante. Il s'agit donc d'un travail de fourmi de visualisation des lignes et de reconstitution grâce à l'intelligence humaine. La situation opposée existe aussi : parfois certaines colonnes étaient dupliquées, mais il y avait des motifs de duplications au sein de la table et il était donc faisable de la nettoyée de façon algorithmique.

L'encodage

Outre le problème du nombre de colonnes, il n'y a pas de choix uniforme d'encodage pour l'ensemble des fichiers. Certains fichiers étaient en ISO et certains en UTF-8. Cependant, pour traiter rapidement toutes les données, il fallait éviter d'avoir à vérifier l'encodage pour ouvrir chaque fichier. Il a donc été décidé de tous les convertir en UTF-8, l'encodage choisi pour le format CoNNL-U que nous avons choisi.

Le token spécial « @ »

Les tokens sont une subdivision de texte à l'échelle qui correspond le mieux au sujet d'étude. Dans notre cas, il s'agit des mots et des signes de ponctuation.

Pour des raisons juridiques, dix tokens consécutifs sont remplacés par des arrobases tous les 200 tokens dans les fichiers textes (.txt) de COHA. [9]

De plus, certains guillemets ont été remplacés par l'expression « ^@ » dans le fichier « sources ». Cependant, il ne s'agit pas d'une décision uniformisée à l'ensemble des titres.

Enfin, des doubles arrobases (« @@ ») sont présentes dans certains ouvrages, notamment dans textID 13 880 et textID 13 879 où elles sont précédées d'un retour à la ligne.

Ces emplois des arrobases engendrent différentes problématiques dans mon travail. Tout d'abord, ma mission est d'étudier le contexte autour des mots pivot de forme « pain ». Or si certains mots sont remplacés par des arrobases, certaines parties du contexte sont perdues ce qui apporte un biais dans mes statistiques.

De plus, ce caractère était censé être caractéristique d'un changement d'ouvrage, notamment lorsqu'il est doublé et précédé d'un retour à la ligne, hors l'utilisation périphérique de ce motif apporte des difficultés lors de l'importation des données : les textes peuvent être tronqués ou interrompre le code, lorsqu'il n'arrive pas à récupérer le textID censé le suivre.

Enfin, lorsque l'arrobase est étudiée d'un point de vue syntaxique ou sémantique, elle est considérée comme une ponctuation. Le remplacement de mots par ce symbole ne fait donc pas que supprimer des informations mais entraîne également des erreurs lors du traitement, qui doivent être rattrapées ensuite.

Le token « @ » n'est pas le seul à être utilisé pour remplacer des mots manquants. Il existe également des tokens remplacés par des points d'interrogation dont le sens ne peut pas être récupéré.

La tokenisation et les espaces dans les fichiers text et wlp

La tokenisation, c'est-à-dire le mode de découpage en tokens, n'est pas parfaitement compatible avec le package spacy couramment employé et notamment dans le corpus de référence francophone de notre projet.

Cette différence de tokenisation peut être dans un premier temps expliqué par des tendances différentes dans des cas limites comme celui des mots composés liés par un ou plusieurs traits d'union. Dans les fichiers wlp, « old-wife » sera un seul token alors que pour spacy il s'agit des tokens « old » et « wife » par exemple. Cependant, tous les problèmes de tokenisation ne sont

pas explicables par des différences de pratiques, il y a également un nombre non négligeable d'erreurs de tokenisation, notamment autour des ponctuations.

Nous avons essayé de régler le problème mais n'aboutissant pas à des résultats concluants, nous avons décidé de faire confiance en la tokenisation des fichiers wlp du Clean COHA [9], entreprise par un laboratoire différent du COHA. Cela a nécessité quelques adaptations mais le résultat fut vraiment très concluant.

Par ailleurs, les ouvrages des fichiers « text » ne sont pas vraiment les textes bruts comme on pourrait s'y attendre : il s'agit en fait de suite de tokens séparés par des espaces. Ainsi, si l'on reprend la phrase précédente on obtiendrait l'exemple ci-dessous. Cette caractéristique n'avait pas d'impact majeur sur notre analyse, nous ne l'avons donc pas modifiée.

« , until clearly and distinctly , " Elnora Comstock , " called the @ @ @ @ @ @ @ @ @ @ One tiny curl added to the top of the first curve of the m in her name , had transformed it from a good »

Figure 5 Exemple d'espaces arbitraires et d'emploi d'arrobases dans un fichier text

Les lemmes manquants dans les fichiers wlp

Alors qu'il est censé y avoir le lemme de chaque mot au sein des tables wlp, certains sont des « colonnes manquantes » et certains sont définie avec des balises qui provoquent des erreurs [9]. Cependant, il est possible de retrouver le lemme grâce à spacy, c'est donc ce que j'ai décidé de faire.

Les auteurs, les titres, les genres et les nombres de mots

Les auteurs

Même si les auteurs sont censés être présents dans COHA, ce n'est pas la réalité du corpus. Dans la majorité des cas, les métadonnées ne contiennent pas les auteurs des ouvrages, et lorsqu'ils sont présents, ils ne sont pas classés dans une colonne particulière. Pourtant, il était important de savoir s'il existait un ou plusieurs auteurs prédominants, notamment dans le genre fiction. Cela pourrait produire un biais tel que, sur sa période, les statistiques seraient davantage représentatives de cet auteur plutôt que de son époque. Etant donné que ces textes de fiction sont majoritairement issus du projet Gutemberg, il a donc été décidé d'utiliser Gutendex, API de ce projet pour essayer de retrouver les auteurs. Il a été également envisagé de recourir à l'API de Wikipédia anglophone avec, pour entrées, les titres des ouvrages.

L'API de Wikipédia étant plus facile à prendre en mains, c'est donc celle-ci sur laquelle j'ai travaillée avec Yoann Dupont (voir annexe). Le principe est simple : nous demandons un titre à l'API, qui cherche un article homonyme dans sa base de données, puis nous cherchons dans les informations renvoyées la caractéristique « auteur » quand elle est présente. Cependant, si un article à le même nom qu'un de nos ouvrages, notre roman peut être reconnu comme une chanson, un film, une région française ou encore un matériau métallique. Cette solution est donc approximative mais permet de dégrossir l'annotation manuelle qui sera de toute façon nécessaire.

Les titres

Plusieurs titres contenus dans la table « sources » ont été identifiés comme faux. C'est en particulier le cas des titres « XXX_YYY » et « poetry ».

Dans le premier cas, après des essais infructueux, je me suis tournée vers Mathieu Dehouck, chargé de recherche au Lattice, qui a eu l'idée de retrouver les noms des revues grâce aux

pieds de pages présents dans les textes. Cette solution n'est pas miraculeuse mais a tout de même permis de récupérer les métadonnées de trois des six identifiants, comme montré ci-dessous.

Dans le second cas, les ouvrages titrés « poetry » n'ont pas la bonne organisation de données : l'auteur, la date et le titre de l'ouvrage sont rassemblés dans la même cellule. Cependant, même si leur formalisme diffère de l'organisation pour le reste du corpus, ils ont entre eux la même organisation, ce qui permet de les reformater.

textID	yearPub	title	fichierText
10792	1931	XXX YYY	text_fic_1930.txt
10793	1931	XXX YYY	text_fic_1930.txt
10794	1931	Red pepper returns	text_fic_1930.txt
10795	1931	Good Pals	text_fic_1930.txt
10796	1931	The lively lady	text_fic_1930.txt
10797	1931	XXX YYY	text_fic_1930.txt

Figure 6 Titres manquants "XXX YYY" dans le fichier sources

Les genres

Même si les genres sont précisés dans le nom du fichier texte contenant le textID, de nombreuses erreurs ont été relevées lors du traitement des données, particulièrement entre les genres fiction et académique. Des articles issus de la revue Plos One sont par exemple considérés comme de la fiction alors que des nouvelles de Charles Dickens seraient des articles académiques.

Face à la détection de plus en plus fréquente d'erreur de classification, plusieurs méthodes ont été envisagés consécutivement. Ayant pour priorité la classification médical / non médical, j'ai tout d'abord pensé à utiliser la base de données de listofjournals.com qui permet une correspondance entre le nom des revues et les thématiques de recherches. Les noms des revues apparaissant dans les titres, je me suis dit naïvement que cela allait correspondre. Je n'ai obtenu aucun résultat, et ai donc dû réfléchir à d'autres options. J'ai ensuite pensé à une classification par mot clés anatomiques et neurologiques, mais Astrid Chevance m'a expliquée que c'était davantage l'association entre certains termes techniques que les termes techniques eux-mêmes qui me permettrait de les reconnaître. Nous avons également pensé avec Mathieu et Yoann passé par des détecteurs de plagiat pour récupérer le nom de la revue, cependant, une faible proportion de textes étaient détectés (les plus récents) et nous obtenions un lien url et non un nom de revue directement.

Nous pensions que nous devions renoncer à la classification de nos œuvres, mais c'était sans compter sur le fait que l'on fasse une découverte par sérendipité. En effet, en cherchant les auteurs dans Wikidata, nous nous sommes rendus comptes avec Yoann qu'il existait un QID qui classifiait l'article dans une multitude de catégories. Dans certains cas, ces noms de catégorie correspondaient à des types d'ouvrages ou bien à des styles littéraires. J'ai donc pu faire un parallèle entre certaines étiquettes de Wikidata et certaines côtes Clément, système de classification des bibliothèques, et notamment celui de la BnF.

Les nombres de mots

Il est également à noter que les nombres de mots indiqués pour chaque ouvrage dans la table « sources » ne correspondent à aucune réalité mesurée. Ce n'est pas une donnée que nous utilisons pour notre étude, mais cela s'ajoute aux nombreux problèmes de COHA.

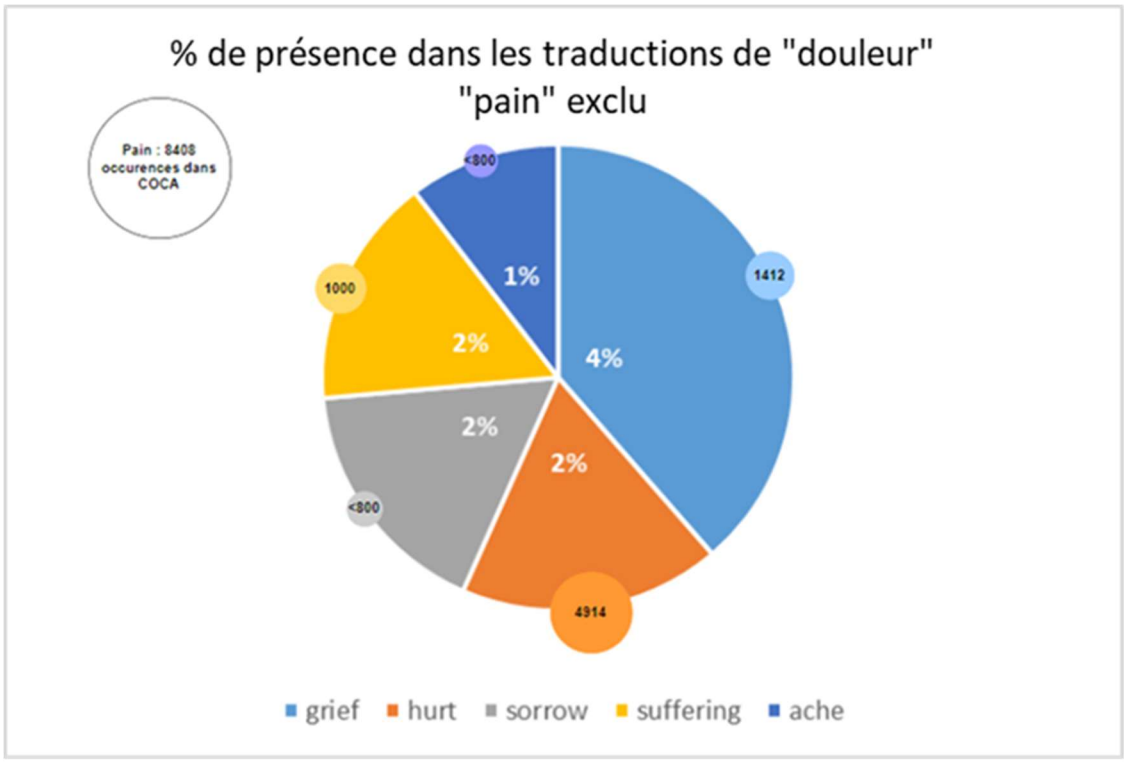
Les langues dans le corpus

COHA est un corpus historique anglophone ; il ne devrait contenir que des textes anglophones. Cependant, j'ai quand-même trouvé par hasard le textID 4166104 hispanique. J'ai donc entrepris un test systématique des titres et contenus des ouvrages grâce à la librairie LanguageIdentifier sur les titres et des extraits de texte. Elle est particulièrement adaptée à cette tâche puisqu'elle a été entraînée sur des textes courts et a donc un meilleur résultat sur des suites de mots de cette taille. Cependant, cela demande une vérification manuelle puisque les mots clés et les « @ » et « ? » remplaçant des mots tronqués ne sont pas reconnus comme de l'anglais. Lorsqu'ils sont très majoritaires dans les suites de mots testés, le résultat est alors faussé.

3. Méthodes

Dans le cadre de mon stage, je suis chargée de construire un corpus anglophone comparable au corpus francophone en me concentrant sur la période 1820-2010 grâce au corpus COHA.

Afin d'étudier la douleur, il a été retenu comme pivots (ou termes centraux de l'analyse) « douleur » et « douloureux », pour leur proximité syntaxique et leurs usages longitudinaux à travers les siècles. Il a ensuite été choisi les termes de « pain » et « painful » pour l'anglais, « pain » étant la traduction dans 80% des contextes du terme « douleur » même si d'autres traductions sont possibles, comme montré ci-dessous.



https://www,kilgarriff,co,uk/bnc-readme,html#lemmatised
https://treq,korpus,cz/index,phpc

Figure 7. Nombre d'occurrences dans COCA (version contemporaine de COHA) et pourcentages de contextes comportant « douleur » en français, traduit par d'autres termes que « pain » en anglais.

Je dois alors m'appuyer sur ses pivots pour répondre aux problématiques suivantes afin que puissent être comparé les usages liés à la douleur en anglais et en français.

- Quelle est la répartition entre 'pain' et 'painful' au sein des pivots, par genre, au fil du temps ?
- Quelle est la fréquence des verbes au singulier et au pluriel dans les phrases avec au moins un pivot, par genre, au fil du temps ?
- Quelle est la fréquence des déterminants au singulier ou au pluriel rattachés directement au pivot, par genre, au fil du temps ?
- Quelles sont les fréquences des expressions ou mots dans les phrases avec au moins un pivot, par genre, au fil du temps ?

Mot ou expression en français	**Mot ou expression en anglais**
Douleur morale	Moral pain
Douleur mentale	Mental pain
Psychalgie (synonyme de douleur psychologique)	Psychalgia
Douleur psychologique	Psychological pain
Douleur psychique	Psychic pain
Douleur physique	Physical pain
Douleur somatique	Somatic pain
Folie	Madness

Aliénation	Alienation
Dépression	Depression
Mélancolie	Melancholy
Spleen	Spleen
Manie	Mania
Délire	Delirium, delusion
Hallucination	Hallucination
Psychose	Psychose
Suicide	Suicide

Figure 8. Tables de correspondances entre mots et expressions cherchées en français et en anglais

- Dans quelles proportions les verbes dans les phrases contenant des pivots sont conjugués à la première personne ou à la troisième personne du singulier, par genre, au fil du temps ?
- Quelles sont les fréquences de parties du corps associées directement à un pivot, par genre, au fil du temps ?
- Quelle est la répartition entre phrases contenant des pivots ayant une tonalité émotionnelle négative et celles ayant une tonalité émotionnelle positive, par genre, au fil du temps ?

Il est ensuite prévu que mon laboratoire s'appuie ultérieurement sur ses premières explorations, après distinction des exemples relevant de la douleur physique et la douleur mentale, afin de répondre aux questionnements globaux ci-dessous.

1. Il existe une simultanéité entre l'apparition de la distinction entre douleur mentale et douleur physique et le développement de la médecine anatomo-pathologique, l'utilisation du premier analgésique pour réduire la douleur physique et l'émergence de la médecine morale (ancêtre de la psychiatrie actuelle) entre 1780 et 1839.
2. Il existe une simultanéité entre une augmentation proportionnelle de l'usage du terme de « douleur » pour parler de douleur mentale et l'utilisation de ce terme par plusieurs aliénistes pour qualifier le symptôme principal de troubles mentaux entre 1840 et la veille de la première guerre mondiale en 1909.
3. Il existe une simultanéité entre une disparition progressive du concept de douleur mentale et l'avènement de la première guerre mondiale entre 1910 et 2010.

4. Résultats et discussions

4.1 Analyse des résultats

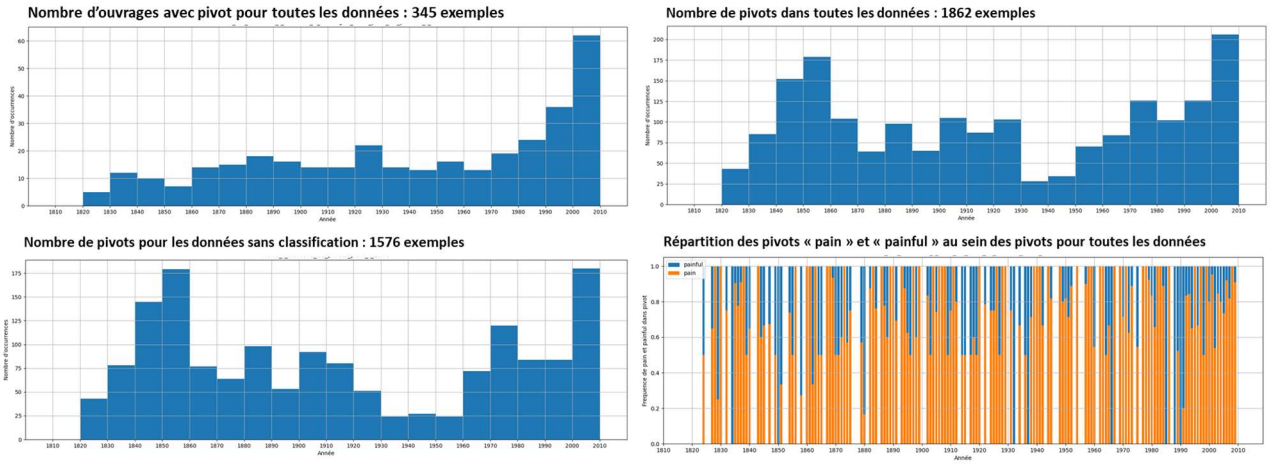


Figure 9. Présentation des données en termes de nombre d'ouvrage avec un pivot, de nombre d'occurrence de pivot et de répartition entre les deux pivots au fil des décénies

Nous pouvons voir ci-dessus que le nombre d'ouvrages contenant des pivots par année pour l'ensemble des catégories d'ouvrages est globalement croissant entre 1820 et 2010 pourtant, lorsque l'on regarde le nombre de pivots par années, il a deux pics de croissance majoritaires avec des sommets locaux atteints dans les années 1850 et 2000. Ainsi, plus d'ouvrages ne veut pas forcément dire plus de pivots. Cette caractéristique a aussi été relevée pour le corpus issu de Gallica pour la version francophone et il a été décidé de normaliser les fréquences par période pour le biais de la disparité des occurrences.

J'ai également voulu faire des analyses par catégorie d'ouvrage, seulement, plus de 80 % des ouvrages n'ont pas pu être catégorisés par mon algorithme. Il ne faut donc pas s'étonner si la majorité des résultats se concentre dans cette « catégorie » d'ouvrages.

Par ailleurs, j'ai entrepris un graphique de la répartition entre « pain » et « painful » au sein des pivots, par année, au fil du temps. « pain » est majoritaire dans nos exemples lorsque l'on prend l'ensemble des données, comme montré ci-dessus.

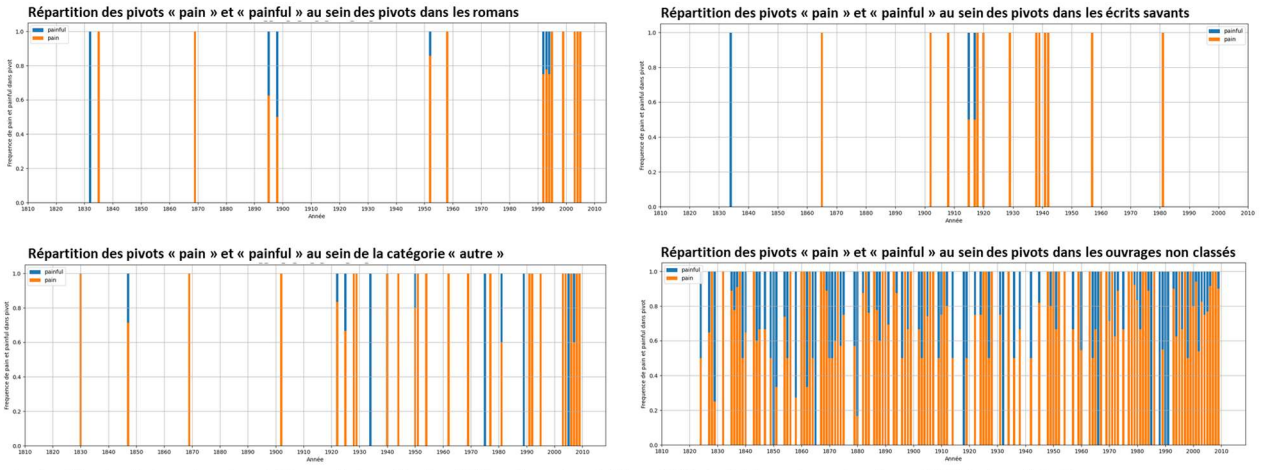


Figure 10 Répartition entre les deux pivots par genre au fil des siècles

De plus, comme on peut le remarquer sur cette nouvelle figure, cette sur-représentativité de « pain » est généralisable dans toutes les sous-ensembles, même si le résultat est moins manifeste lorsqu'il y a peu de décennies avec des données.

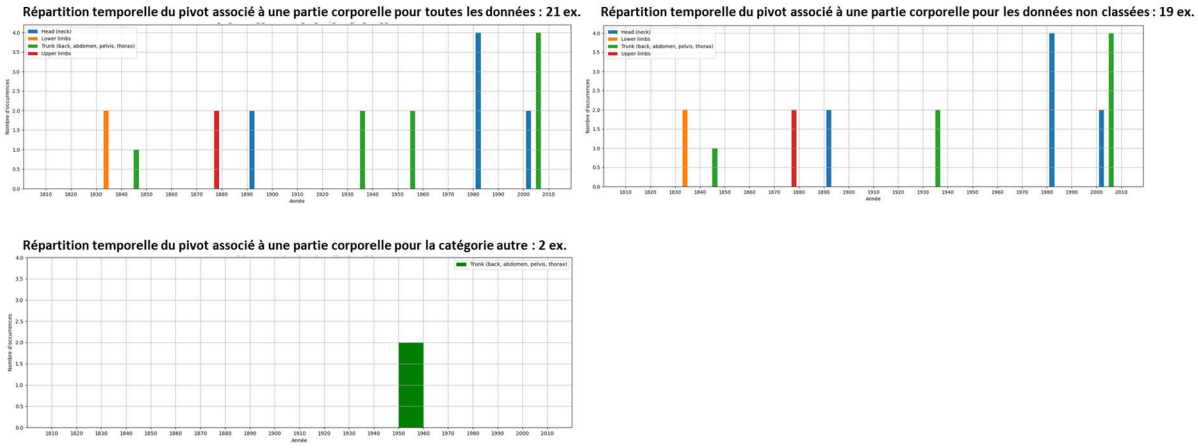


Figure 11. Occurrences de parties associées à des zones corporelles pour tous les ouvrages et par genre au fil des siècles

Ensuite, conformément à la volonté d'obtenir des corpus comparables, j'ai traduit les termes scientifiques relatifs aux parties du corps utilisés dans le corpus français regroupés avec la même classification en zones employées en première approximation. J'ai ensuite fait des analyses d'occurrence de termes associés à ces zones, rattachés directement sémantiquement à un pivot. Seulement, comme le montre la figure ci-dessus, le nombre d'occurrences est vraiment faible et cette approche sémantique serait difficilement exploitable. Des pistes sont à exploiter du côté des bases de données TA2 et TNA de la FIPAT deux bases de données de termes anatomiques classés par zones corporelles. Cependant, le corpus central étant le corpus francophone, cette piste ne sera probablement pas exploitée.

Roman

Classifieur	Date	Quantité
Physical pain	1869	1

Ecrits savants

Classifieur	Date	Quantité
physical pain	1941	1
depression	1915	1

Théologie

Classifieur	Date	Quantité
delusion	2001	1

Autre

Classifieur	Date	Quantité
physical pain	1922	1
physical pain	1950	1

None

Classifieur	Date	Quantité
physical pain	1959	2
physical pain	1994	2
physical pain	1995	3
physical pain	2006	2
alienation	1891	1
delirium	1864	2
depression	1905	2
depression	2001	2
madness	1890	2
melancholy	1828	1
melancholy	1847	2
melancholy	1890	2
suicide	1864	2

Figure 12. Tableau des seules occurrences comportant les mots et expressions recherchées par genre d'ouvrage

Dans la même dynamique, j'ai cherché les mots ou expressions imposées par le protocole dans l'espoir qu'ils puissent servir comme pré-classification entre douleur mentale et douleur physique, afin de soulager l'annotation manuelle indispensable à la classification par réseau de neurones. Seulement, les occurrences sont une nouvelle fois très faibles et ne sont donc pas exploitables, tant dans une

optique d'analyse syntaxique que dans une optique d'aide à l'annotation, comme le montre les tables des exemples exhaustifs obtenus.

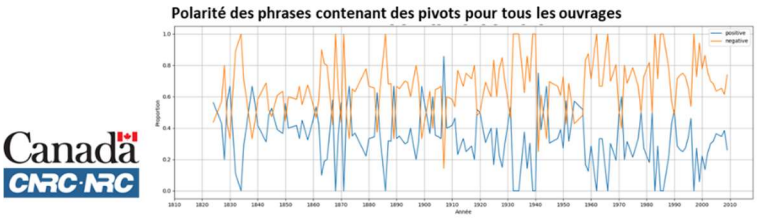


Figure 13. Polarité émotionnelle des phrases contenant un pivot au fil des siècles

J'ai ensuite tracé l'évolution de la polarité émotionnelle des phrases contenant au moins un pivot au fil des siècles grâce à la base de données du CNRC-NRC, comme montré ci-dessus. La polarité majoritaire est la polarité négative, même s'il existe des inversions de tendance sur des courtes périodes. Ce résultat n'est pas ce qui était attendu puisqu'en l'état des lieux, la même analyse sur le corpus Gallica avec l'outil francophone dérivé de l'outil du CNRC-NRC montre plutôt une majorité de polarité positive associée probablement à une romantisation de la douleur. Il s'agit peut-être d'une différence culturo-linguistique entre les deux langues. Cependant, le corpus français étant bien plus important en terme de nombre d'exemples, il serait aussi important d'écarter la possibilité que notre échantillon ne soit pas représentatif.

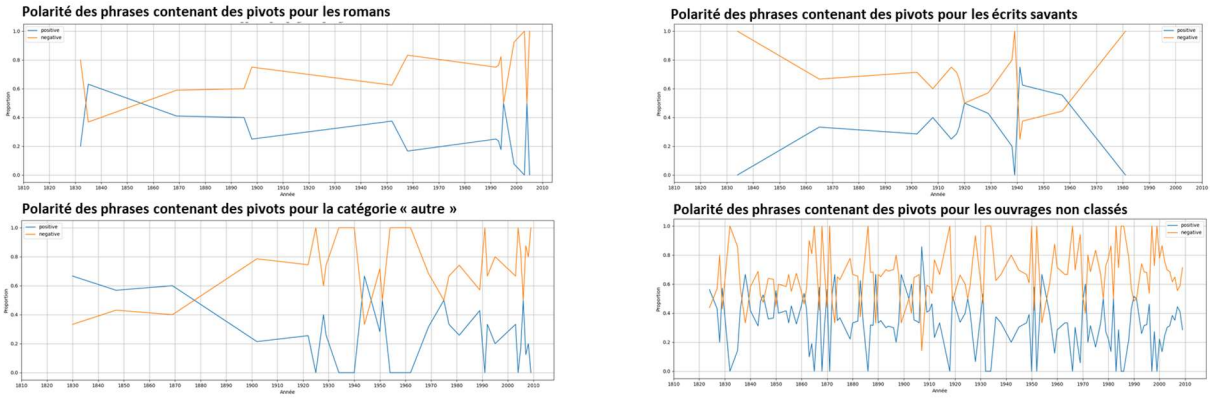


Figure 14. Polarité émotionnelle des phrases contenant un pivot au fil des siècles par genre

L'analyse de la polarité a également été faite pour toutes les catégories d'ouvrages qui le permettait, comme présenté ci-dessus et la tendance à une majorité de phrases avec pain et une émotion négative se confirme. L'inversion de polarité entre les années 1940 et 1980 pour les écrits savants pourrait être très intéressante à approfondir. Il faudrait attendre une meilleure classification des ouvrages pour infirmer ou affirmer cette propriété.

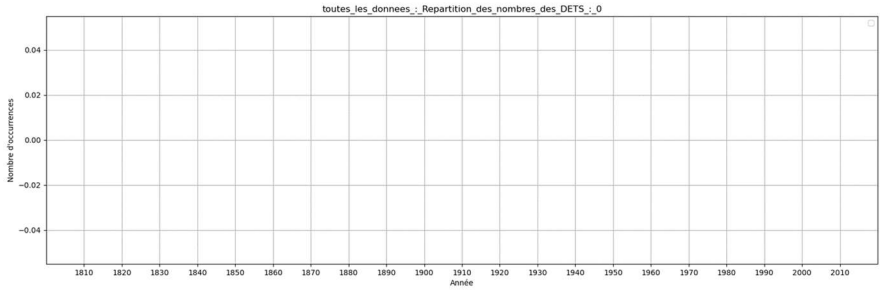


Figure 15. Tentative échouée d'analyse de déterminants qualifiants un pivot au fil des siècles

Je suis ensuite passée à l'analyse plus syntaxique des phrases contenant un pivot et je me suis premièrement attardée sur les déterminants rattachés à un pivot, dans le but de savoir s'ils étaient plus au pluriel ou au singulier.

Cependant, comme vous pouvez le voir sur le graphique ci-dessus, mon algorithme n'en a trouvé aucun. Plusieurs pistes sont envisagées. La plus probable est que le modèle d'annotation automatique de la librairie Spacy que nous avons utilisé est défectueux dans les phrases contenant au moins un pivot. C'est une possibilité non négligeable, même si les modèles anglophones sont censés être beaucoup plus fiables que les modèles francophones équivalents par exemple.

La deuxième possibilité, plus humble, mais qui n'est pas la plus envisagée au sein de l'équipe, est qu'il y a une erreur de filtrage de donnée au sein de mon algorithme qui toucherait l'ensemble des exemples avec des déterminants, même sans que l'erreur soit spécifiquement ciblée sur ces exemples. Cependant, de nombreux tests de vérification ont été faits à toutes les étapes du traitement et il serait étonnant qu'une perte massive soit encore présente.

Enfin, dernière possibilité, qui ne pourra être vérifiée que par preuve qu'il n'y a aucune erreur, peut-être qu'aucune donnée de correspond à cette configuration syntaxique, aussi improbable que ça puisse l'être.

Figure 16 Répartition des emplois de la première et la troisième personne comme sujet de verbes contenus dans des phrases avec un pivot pour tous les ouvrages puis par genre

Enfin, la dernière analyse correspond plus ou moins à la précédente mais porte sur les verbes. Nous avons voulu chercher si le sujet du verbe associé avec pain était plutôt associé à une première personne du singulier ou à une troisième personne du singulier. Dans une première approximation, nous avons alors voulu savoir à quelle personne était accordé les verbes présents dans les mêmes phrases que les pivots. Comme le montre les graphiques ci-dessus, la structure majoritaire est celle de l'auxiliaire avec une troisième personne, suivie par celle du verbe à la troisième personne. L'emploi de la première personne est extrêmement minoritaire et toujours associé à un auxiliaire. Ce résultat peut encore une fois être dû à des problèmes d'annotation syntaxiques des données. Cependant, si ce n'est pas le cas, il faudrait séparer les verbes ayant un pivot comme sujet (travail déjà fait mais non exploité pour le moment), des verbes où un pivot serait complément, et ensuite faire l'analyse sur cette deuxième catégorie. Il serait également intéressant d'exploiter les différentes modalités anglophones ayant une grande coloration sémantique afin de mieux comprendre pourquoi l'emploi de la première personne est uniquement associé à un auxiliaire.

4.2 Voies d'approfondissement

En conclusion, même si mon travail ne me permet pas de valider ou non des hypothèses de recherche puisqu'il s'agit d'un travail exploratoire, la piste de la polarité émotionnelle semble être une bonne piste à explorer. Cependant, aucune de mes analyses ne permet pour l'instant de trouver des hypothèses de corrélation entre les changements de polarité et des caractéristiques sémantique ou syntaxiques indiscutables. Il serait également important de poursuivre la recherche d'autres corpus pour augmenter le nombre d'exemple, notamment pour le 19ème siècle, commencée par le LATTICE avant ma venue.

Cependant, d'autres tâches parallèles m'attendent. Je suis par exemple en train de finir une classification sémantique et francophone d'adjectifs liés à la douleur avec Alexia (voir en annexe). Ceci aurait pour but à court terme de comparer les usages en français et en anglais et également de faire une étude longitudinale de ses différentes catégories, si le nombre d'occurrences le permet. A plus long terme, elle pourrait être utilisée comme base pour un outil d'aide à l'objectivation du niveau et

du type de douleur ressentie par le patient afin de mieux le prendre en charge, notamment pour les douleurs psychiques.

En parallèle, il est prévu que j'entreprenne un lexique associé aux problèmes de santés physique et psychologique grâce à la classification des articles de Wikipédia, déjà étiquetés, pour certains, dans ces deux catégories.

Enfin, dans le corpus Gallica, les articles sont étiquetés entre « médical » et « non médical » en plus de leur classification en genre (Roman, Théologie…) puisque la classification documentaliste de la BnF le permettait. Ce n'est malheureusement pas le cas pour nos données, mais il est envisagé que je sois aidée par Qi et Frédérique (voir en annexes) pour adapter un réseau de neurones afin de classer les documents dans ces deux catégories.

5. Bibliographie

[1] Ducasse, D. et al. (2018). Psychological Pain in Suicidality: A Meta-Analysis. J. Clin. Psychiatry, (79).

Verrocchio, M. C. et al. (2016). Mental Pain and Suicide: A Systematic Review of the Literature. Front. Psychiatry, (7).

[2] Chevance, A. et al. (2020). Identifying outcomes for depression that matter to patients, informal caregivers, and health-care professionals: qualitative content analysis of a large international online survey. Lancet Psychiatry, (7).

[3] Charvet, C. et al. (2022). How to measure mental pain: a systematic review assessing measures of mental pain. Evid. Based Ment. Health.

[4] UNIVERSAL DEPENDENCIES. CoNLL-U Format. [en ligne] (page consultée le 30/05/2024) https://universaldependencies.org/format.html

[5] DAS LEIBNIZ-INSITUT FÜR DEUTSCHE SPRACHE. Workshop on the Challenges in the Management of Large Corpora [en ligne] (page consultée le 30/05/2024) https://corpora.ids-mannheim.de/cmlc.html

[6] LANGUAGE TECHNOLOGY GROUP – UNIVERSITY OF EDINBURGH – THOMPSON. Corpus Creation for Data-Intensive Linguistics [en ligne] (page consultée le 30/05/2024) https://www.cogsci.ed.ac.uk/~ht/harold.html

[7] INTERACTIONS, CORPUS, APPRENTISSAGES, REPRESENTATIONS – CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE - WYNNE. AHDS Guides to Good Pratice, Developing Linguistic Corpora : a Guide to Good Practice [en ligne] (page consultée le 30/05/2024) https://icar.cnrs.fr//ecole_thematique/contaci/documents/Baude/wynne.pdf

[8] ENGLISH CORPORA. . [en ligne] (page consultée le 30/05/2024) https://www.english-corpora.org/siteLicense-fulltext.asp

[9] Alatrash, R. Schlechtweg, D. Kuhn, J. et al. (2020). CCOHA: Clean Corpus of Historical American English. LREC, (page consultée le 30/05/2024) https://aclanthology.org/2020.lrec-1.859/

[10] ENGLISH CORPORA. Corpus of Historical American English [en ligne] (page consultée le 30/05/2024) https://www.english-corpora.org/coha/

[11] CORPUS DATA. Full-text corpus data [en ligne] (page consultée le 30/05/2024) https://www.corpusdata.org/formats.asp

[13] CORPUS DATA. Full-text copus data from corpus.byu.edu [en ligne] (page consultée le 30/05/2024) https://web.archive.org/web/20170409215611/https://www.corpusdata.org/database.asp/

[14] NATCORP. British National Corpus [en ligne] (page consultée le 30/05/2024) http://www.natcorp.ox.ac.uk/

NATCORP, LEECH. A brief users's guide to the grammatical tagging of the British National Corpus [en ligne] (page consultée le 30/05/2024) http://www.natcorp.ox.ac.uk/docs/gramtag.html

LANCASTER UNIVERSITY. BNC2 POS-tagging Manual guidelines to wordclass tagging [en ligne] (page consultée le 30/05/2024) https://ucrel.lancs.ac.uk/bnc2/bnc2guide.htm

[15] OFFICE QUEBECOIS DE LA LANGUE FRANCAISE. Grand dictionnaire terminologique lemme [en ligne] (page consultée le 30/05/2024) https://vitrinelinguistique.oqlf.gouv.qc.ca/fiche-gdt/fiche/26570748/lemme

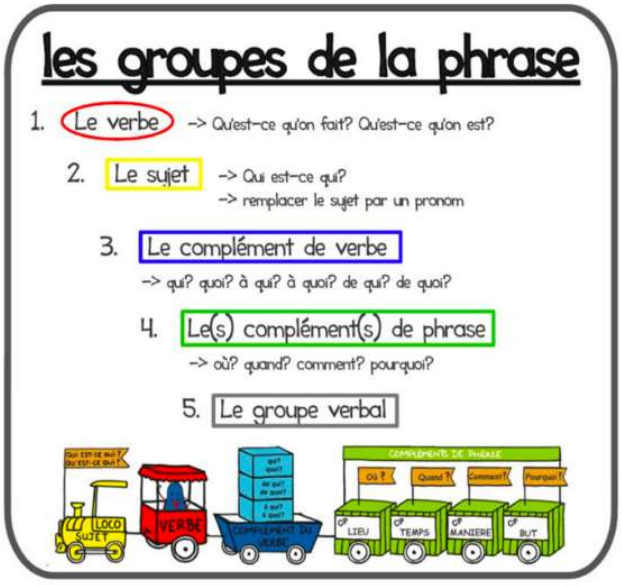
6. Annexes

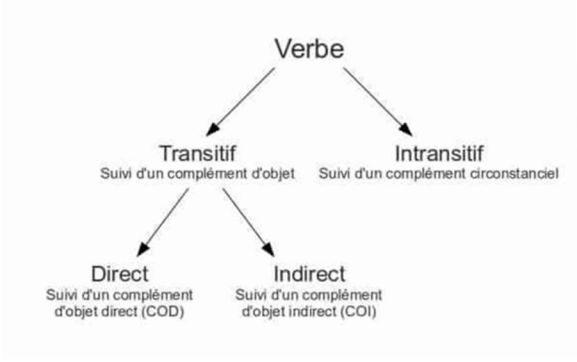
Je tenais à souligner la cohérence de ce stage dans mon parcours académique et professionnel. En effet, Charpak m'a permis d'avoir des connaissances dans de nombreux domaines scientifiques qui me permettent de comprendre n'importe quel type de données et de savoir les traiter informatiquement. Cependant, ce que Charpak ne pouvait pas m'apporter, je l'ai trouvé dans mes deux stages effectués en licence.

J'ai tout d'abord appris à manipuler des données médicales très protégées et à effectuer des analyses émotionnelles au sein du laboratoire NLP l'année dernière. C'était une première entrée dans la matière, mais certainement suffisant pour réussir à être une future chercheuse aboutie.

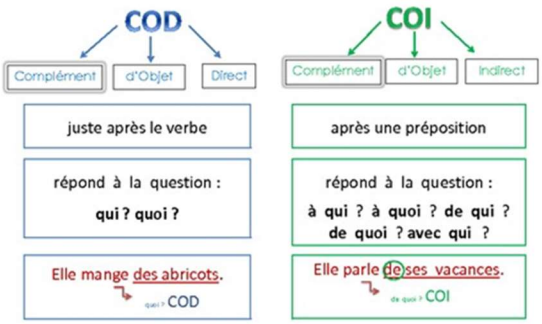
C'est pourquoi, j'ai poursuivi mon chemin en m'arrêtant cette année au LATTICE ce qui me permet de développer des compétences en élaboration de corpus et de savoir comment structurer des données afin d'étudier des notions particulières comme la douleur ici, ou le manque pour Bibou (après mon déménagement et mon inscription dans ma nouvelle fac). De plus, après ma soutenance il est prévu que je sois formée au LATTICE en terme de réseaux de neurones. Cela me permettra d'avoir de meilleurs arguments pour approfondir cette approche que j'ai longtemps refusée mais qui reste tout de même indispensable pour certaines tâches et surtout pour mon mémoire et me faire employer.

Je pense donc avoir fait des choix en parfaite adéquation avec le profil de chercheur en traitement automatique du langage qui sauront complétés mon master dans ce même domaine l'année prochaine.

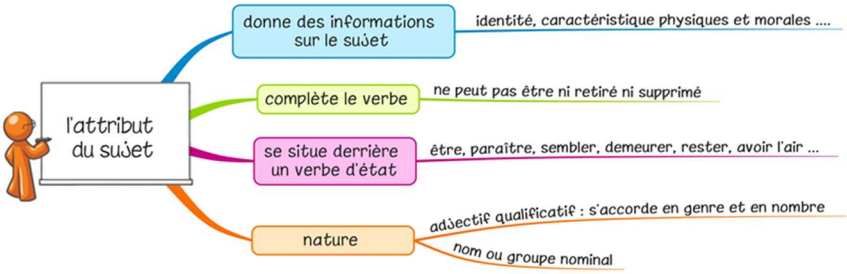




Verbes transitifs



Verbes attributifs



Verbes intransitifs