Insertion des catégories grammaticales dans le Dictionnaire Électronique des Synonymes (DES) -Document de travail

Laurette Chardon

Insertion des catégories grammaticales dans le Dictionnaire Électronique des Synonymes (DES) - Document de travail

Laurette Chardon

January 25, 2023

Abstract

Ce document détaille la procédure suivie pour introduire la catégorie grammaticale dans les 50.373 entrées du Dictionnaire Électronique des Synonymes (DES) [1] du laboratoire CRISCO [2]. **Ce travail a eu lieu de janvier 2021 à novembre 2022**. Il s'est déroulé en deux grandes phases :

- à partir des données de l'ATILF (*Analyse et Traitement Informatique de la Langue Française, UMR7118*), laboratoire du CNRS [3] chargé de la maintenance et du développement du TLFi (*Trésor de la Langue Française informatisé*), transmises sous forme d'un classeur (au format .xlsx) contenant les catégories grammaticales. Avec des programme en Python, les verbes, substantifs, adjectifs et adverbes ont été introduits dans la base de données Mysql du DES. Ensuite une recherche par schémas (*-ais, -euse, -ale, -ande,...*) avec plusieurs groupes de mots mélangés, suivie d'un traitement automatique ont permis d'exploiter la totalité du fichier d'origine.

- à partir d'autres moyens : avec la librairie Spacy en langage Python, par recherche de schémas avec un tableur et en complétant manuellement. [4]

Ce document de travail est accompagné d'un dépôt git public : https://git.unicaen.fr/crisco-des-public/descatgram

Contents

[1] https://crisco.unicaen.fr/des/
[2] https://crisco.unicaen.fr
[3] https://www.atilf.fr/
[4] Un grand merci à Jacques François, professeur associé au CRISCO, pour son aide

1 Introduction

L'idée d'introduire les catégories grammaticales dans le DES n'est pas récente. Elle s'est naturellement imposée à l'esprit des responsables du DES et, en particulier, suite aux retours de plusieurs internautes fidèles[5].

Nous sommes donc partis des données de l'ATILF correspondant aux informations affichées sur https://www.cnrtl.fr/definition et avons défini une procédure de traitement que nous exposons dans ce rapport.

2 A partir du fichier des lemmes du TLFi

2.1 Etude préliminaire du fichier

Ce fichier intitulé *TLFI complet lemmes.xls* se présente sous la forme d'un classeur avec 54.280 lignes (entrées) lisible avec un tableur. Un premier examen manuel nous montre que les lemmes apparaissent dans la première colonne. Lorsqu'il y a plusieurs entrées dégroupées dues, entre autres à des origines étymologiques différentes, le lemme se termine par un chiffre de 1 à 6.

Apparaissent ensuite dans les colonnes suivantes :

- un complément : par exemple *-ée* pour les participes passés (*accosté, accouché, mouvementé,...*),
- la forme féminine pour des adjectifs ou des substantifs : *-euse, -ante, -ienne, -ète* , par exemple, *mousseux,-euse, migrant,-ante, milicien,-ienne, discret,-ète, ...*
- une autre forme orthographique
- la catégorie grammaticale

La table 1 nous en donne un extrait avec la majorité des différents cas de figure.

Nous voyons que le fichier contient plusieurs lignes hétérogènes. On peut arriver ainsi jusqu'à un maximum de 7 colonnes (voir table 1)

Un traitement avec un programme en Python nous donne les informations suivantes :

- 2.274 lignes sur lesquelles la colonne 1 se termine par 1 (comme MEUBLE1 dans la table 2)
- 2.281 lignes sur lesquelles la colonne 1 se termine par 2
- 297 lignes sur lesquelles la colonne 1 se termine par 3
- 44 lignes sur lesquelles la colonne 1 se termine par 4
- 7 lignes sur lesquelles la colonne 1 se termine par 5
- 1 lignes sur lesquelles la colonne 1 se termine par 6
- aucune ligne avec la colonne 1 se terminant par 7

Le calcul est réalisé sur la première colonne uniquement, celle qui donne le lemme.

Panne est le lemme avec le plus d'acceptions (6) toutes en tant que substantif féminin. Les six lemmes avec cinq acceptions sont : *pique, pointer, baba, canette, coco* et *faire*.

Le programme a également calculé le nombre d'entrées pour les catégories grammaticales suivantes présentes dans la seconde colonne (celle qui est la mieux renseignée par les catégories grammaticales) :

- 6.977 verbes
- 30.209 substantifs
- 5.181 adjectifs
- 1.115 adverbes

[5] Voir les commentaires sur cet article dans le blog de la MRSH de Caen : https://mrsh.hypotheses.org/5578#comment-334

Col 1	Col 2	Col3	Col4	Col5
ABATTRE	ABBATTRE	ABATRE	verbe trans.	
ABBATTRE	voir ABATTRE			
ABATTU	UE			
...				
ABORAL	ALE	AUX	adj.	
...				
ABOTÉ	ÉE	ABOTTÉ	ÉE	adj.
...				
AUTO(-)DESTRUCTEUR	TRICE	AUTO(-) DESTRUCTIF	IVE	adj.
...				
AUTOPORTANT	ANTE	AUTOPORTEUR	EUSE	adj. et subst.
...				
MÉTROPOLITAIN1	-AINE	adj.		
MÉTROPOLITAIN2	-AINE	subst. masc. et adj.	MÉTRO	subst. masc.
MÉTROPOLITE	subst. masc.			
METS	subst. masc.			
METTABLE	adj.			
METTEUR	-EUSE	subst.		
METTON	subst. masc.			
METTRE	verbe			
MIS MISE	part. passé et adj.			
MÉTURE	subst. fém.			
MEUBLANT	-ANTE	part. prés. et adj.		
MEUBLE1	subst. masc.			
MEUBLE2	adj.			
MEUBLE3	adj. et subst.			

Table 1: Extrait du fichier de départ du TLFi

ASSESSORAL	ALE	AUX	ASSESSORIAL	IALE	IAUX	adj.

Table 2: Exemple avec un lemme sur 7 colonnes

Par cette méthode, nous avons juste essayé d'extraire quelques informations rapidement sans chercher à être rigoureux dans le traitement. Par exemple, un lemme considéré comme verbe dans la colonne 2 peut-être un susbstantif dans la colonne 3 et ainsi de suite. Les chiffres donnés représentent donc la marge basse puisque nous n'avons tenu compte que des catégories grammaticales dans la seconde colonne uniquement.

2.2 Introduire les verbes

Le fichier *TLFI complet lemmes.xls* contient un onglet *Verbes* que nous avons utilisé et sauvegardé dans un autre fichier CSV (Comma Separated Values) recommandé pour l'interopérabilité des données.

Un extrait de ce dernier est donné dans la table 3.

Col 1	Col 2	Col 3
ABADER (S')	verbe pronom.	ABADER
ABAISSER	verbe trans.	ABAISS
ABALOBER	verbe trans.	ABALOB
ABALOURDIR	verbe trans.	ABALOU
ABANDONNER	verbe trans.	ABANDO
ABASOURDIR	verbe trans.	ABASOU
ABÂTARDIR	verbe.	ABÂTAR
ABCÉDER	verbe intrans.	ABCÉDE
ABDIQUER	verbe trans.	ABDIQU
ABEAUDIR (S')	verbe pronom.	ABEAUD
ABERRER	verbe intrans.	ABERRE
...		
ACCIDENTER1	verbe trans.	ACCIDE
ACCIDENTER2	verbe trans.	ACCIDE

Table 3: Extrait de l'onglet Verbes du fichier de départ

Nous nous sommes servis des deux premières colonnes.
Dans la première colonne, tout apparait en majuscule : il fallait donc transformer en minuscule et supprimer les espaces au début et à la fin du mot pour retrouver l'entrée dans la base de données du DES. Ensuite, les formes pronominales étant indiquées par *s'* ou *se* à la fin du verbe, il fallait les repositionner au début du verbe en minuscule. Toujours dans cette première colonne, des parenthèses ouvrantes et fermantes étaient présentes sur certaines lignes pour indiquer le doublement de consonnes, par exemple *flip(p)er* ou *époumon(n)er*. Ces parenthèses devaient être supprimées. **Enfin si le verbe se terminait par les chiffres de 1 jusqu'à 7, il s'agissait de plusieurs entrées dégroupées. Il fallait donc concaténer les différentes catégories grammaticales des lignes concernées en les séparant par le caractère spécial *point virgule*, caractère spécial choisi pour différencier les acceptions**.
La seconde colonne, concernant la nature grammaticale, peut avoir comme valeurs celles de la table 4.

verbe
verbe trans.
verbe intrans.
verbe pronom.
verbe intrans. et trans.
verbe impers.
verbe trans. indir.
verbe trans. indir. et intrans.
verbe trans. indir. et pronom.

Table 4: Codes grammaticaux traités en tant que verbes

Ces codes ont été conservés tels quels dans le champ *Nature* de la base de données Mysql du DES. Il a été nécessaire cependant de remplacer 492 entrées avec *verbe.* en enlevant le point final dans le tableur.
Un autre traitement réalisé dans le tableur concernait la présence de parenthèses ouvrantes et fermantes dans certaines entrées autres que celles commençant par *s'* ou *se*. Dix entrées ont été

trouvées (voir table 5) et ont donc été traitées manuellement en enlevant les parenthèses et en gardant leurs contenus ou pas selon leur présence dans le DES.

AC(C)OUF(F)LER
AFFOND(R)ER
AP(P)IÉGER
AUTO(-)FÉCONDER
ENTRE(-)DÉCHIRER
ENTRE(-)DÉVORER
ENTRE(-)NUIRE
ENTRE(-)REGARDER
ENTRE(-)SUIVRE
ENTRE(-)TUER

Table 5: Exemples de verbes avec des parenthèses

Dans la base de données Mysql, le champ *Nature* de 20 caractères créé avant cette étude a été allongé à 100 caractères.
Pour ce premier traitement, plusieurs essais ont été nécessaires avant de le finaliser. Il était donc important d'une part de créer un fichier d'enregistrement détaillé des actions (fichier de log) et d'autre part de sauvegarder la base avant traitement et de la restaurer en cas d'erreurs.

Enfin autre point important : si le champ *Nature* dans la base Mysql est déjà renseigné, on vérifie que la nouvelle catégorie à entrer n'est pas déjà insérée. Si c'est le cas, on ne l'ajoute pas. Si ce n'est pas le cas, elle est ajoutée avec un *point virgule*. Le *point virgule* permet donc de reconnaitre les entrées dégroupées ayant des codes grammaticaux différents. Si une entrée dégroupée a la même catégorie grammaticale, cela n'apparaitra pas dans le DES.

Au final 6.982 verbes ont été traités avec 391 verbes dégroupés (se terminant par un chiffre), 1.475 verbes absents du DES et 5.507 verbes mis à jour dans le DES. Le détail des opérations est listé dans le fichier de log du git public du CRISCO [6].

2.3 Introduire les adjectifs

Une copie du fichier TLFI de départ, en enlevant toutes les entrées verbes traitées est créée et nommée *TLFI complet lemmes-20210105.csv*. Dans le tableur, un tri est réalisé pour ne garder que les lignes avec la seconde colonne égale à *adj.*. Nous obtenons ainsi le fichier *TLFI complet lemmes-adj.csv* avec 4.641 lignes.

Un nouveau programme python (InsertAdjBD.py), créé à partir du précédent, réalise les traitements de base (transformer en minuscules et supprimer les espaces superflus au début et à la fin du mot) et la vérification des acceptions (l'adjectif se termine t-il par un chiffre de 1 à 5, nombre maximal d'acceptions repéré ?).

Sur 4.641 lignes traitées, le nombre d'adjectifs considérés comme des entrées dégroupées est de 122, le nombre d'adjectifs absents du DES est de 2.150 et le nombre d'adjectifs présents et modifiés dans le DES est de 2.491.

On peut s'étonner que quasiment la moitié des adjectifs du fichier TLFi ne soit pas présente dans le DES. Si on regarde rapidement le fichier de log [7], on s'aperçoit que 637 d'entre eux se terminent par *-ique*, 175 d'entre eux finissent par *-able* et 34 par *-ible*. Il est vraisemblable que certains ont potentiellement des synonymes et pourraient être insérés dans le DES. Cette étude se focalisant sur l'insertion des catégories grammaticales, la réflexion n'a pas été menée plus loin mais ce fichier de log pourra être une base pour compléter le DES à l'avenir.

Un premier bilan intermédiaire nous permet de conclure que 6.982 verbes et 4.641 adjectifs du

[6] https://git.unicaen.fr/crisco-des-public/descatgram/-/blob/master/CatGramLogInsertionVerbes.csv
[7] Le fichier de log d'insertion des adjectifs est disponible sur le git public du DES : https://git.unicaen.fr/crisco-des-public/descatgram/-/blob/master/CatGramLogInsertAdj.csv

fichier TLFi ont été traités, soit un total de 11.623 sur 54.280 entrées, environ 21%. Dans le DES, 5.507 verbes et 2.491 adjectifs ont été mis à jour sur 50.350, soit 16%.

2.4 Introduire les substantifs

Le traitement des 28.588 entrées traitées n'a pas posé de problèmes particuliers : transformer en minuscule, enlever les espaces superflus au début et à la fin du mot, supprimer les parenthèses sur 533 entrées. Quelques exemples sont données dans la table 6.

AUTO(-)DISCIPLINE
BOUFFON(N)ISTE
CO(-)AUTEUR
CONSON(N)ANCE
COUP(-)DE(-)POING
ESSUIE-MAIN(S)
GARDE-CÔTE(S)
HORS(-)D'OEUVRE
MOYEN(-)ÂGE
TÊTE(-)À(-)TÊTE

Table 6: Extrait de substantifs avec parenthèses dans le fichier TLFi de départ

Concernant la catégorie grammaticale, la table 7 donne les différents cas trouvés. Certains traitements étaient nécessaires pour homogénéiser le champ catégorie grammaticale : trois étaient en majuscules, à transformer en minuscules, des points et des espaces étaient absents pour certaines abréviations ("subst fém." et "subst.fém." à transformer en "subst.(espace)fém." , idem pour subst. masc.).

subst.	subst. et adj.	subst. et adj. fém.
subst. et adj. inv.	subst. et adj. masc.	subst. et interj.
subst. fém.	subst. fém. (plur.)	subst. fém. (plur).
subst. fém. et adj.	subst. fém. et adj. fém.	subst. fém. et adj. inv.
subst. fém. et adv.	subst. fém. et interj.	subst. fém. inv.
subst. fém. ou masc.	subst. fém. plur.	subst. inv.
subst. invar.	subst. masc.	subst. masc. et adj.
subst. masc. et adj. inv.	subst. masc. et adj. masc.	subst. masc. et adv.
subst. masc. et élém. de loc.	subst. masc. et fém.	subst. masc. et fém. plur.
subst. masc. et interj.	subst. masc. et inv.	subst. masc. et loc. adv.
subst. masc. inv.	subst. masc. inv. et adj. inv.	subst. masc. invar.
subst. masc. ou fém.	subst. masc. plur.	subst. masc. plur. et adj.
subst. masc. sing.	subst. masc. sing. inv.	subst. masc. sing. ou fém. sing.
subst. plur.		

Table 7: Codes grammaticaux traités en tant que substantifs

Sur les 28.588 entrées du TLFi traitées, 1.373 étaient dégroupées, 9.445 étaient absents du DES et 19.143 ont été mis à jour dans le DES (Voir fichier de log [8]).

A ce stade, le pourcentage des entrées du fichier TLFi traitées est de 74%.

2.5 Introduire les adverbes

943 adverbes ont été examinés automatiquement (Voir fichier de log [9]). Nous avons un total de 6.982 verbes, 4.641 adjectifs, 28.588 substantifs et 943 adverbes. Ce qui donne un total de 41.154 sur 54.280 entrées du TLFi traitées, soit 75,8%.

Par contre dans le DES, une entrée pouvant à la fois être verbe et/ou substantif et/ou adjectif, les entrées ayant le champ *Nature* renseigné sont au nombre de 25.383 sur 50.350, soit 50%.

[8] Le fichier de log d'insertion des substantifs est disponible sur le git public du DES : https://git.unicaen.fr/crisco-des-public/descatgram/-/blob/master/CatGramLogInsertSubst.csv

[9] https://git.unicaen.fr/crisco-des-public/descatgram/-/blob/master/InsertAdvBD-Log.csv

2.6 Introduire une première catégorie de mots mélangés

A ce stade, nous avons un fichier avec 13.079 entrées TLFi restantes à traiter pour lesquelles la seconde colonne est un mélange de catégories grammaticales (*prép.*pour préposition, *préf.* pour préfixe, *loc.* pour locution, *part. prés.*, ...), de compléments de renvoi (VOIR ABATTRE pour l'entrée ABATRE, voir AFFAMEMENT. pour l'entrée AFFAMATION, ...). Nous devons donc réaliser un travail de vérification et de correction de certaines lignes du fichier TLFi au préalable. C'est la raison pour laquelle nous utilisons le terme de *mots mélangés*. Dans le classeur, nous trions sur cette seconde colonne et nous gardons celles égales à *-acte, -aine, -ainte, -aise, -aite, -ale, -als, -aux, -ande, -ane, -anne, -ante, -apse, -arde, -ate, -aude, -aux, -close, -cuite, -dite, -douce, -dure, -ecte, -ienne, -ée, -éenne, -ées, -elle, -ende, -enne, -ente, -ère, -ète, -ette, -eule, -eure et -euse*.

Sur ces 2.940 lignes retenues, nous avons donc en seconde colonne les extensions ci-dessus et dans les colonnes 3, 4 et 5 les catégories grammaticales qui, après insertion dans le DES, seront séparées par un *point virgule*.

Sur les 2.940 entrées traitées, 62 étaient des acceptions différentes, 852 étaient absentes du DES et 2.088 ont été renseignées dans le DES.

Après traitement un total de 44.094 (41.154 + 2.940) sur 54.280 entrées du TLFi soit 81,2 % ont été utilisées, et dans le DES, 28.155 entrées avec le champ *nature* renseigné sur 50.451 soit 55,8%.

2.7 Introduire une seconde catégorie de mots mélangés

Le même principe exposé dans la précédente section a été suivi : sur la colonne 2, nous gardons celles avec *-ails, -faite, -fine, -haute, -ie, -ielle, -ienne, -ière, -ile, -ille, -incte, -ine, -ique, -ise, -isse, -ite, -ive, -oise, -onne, -onde, -one, -ote, -otte, -oue, -trice, -ue, -une, -use, aine, ainte, aisceau, aise, aisse, aite, ante, arde, aux, ecte, ée, éenne, elle, ente, ère, erse, erte, ète, ette, euse, ie, ienne, oise, onne, trice*.

Nous retenons 5.010 entrées qui une fois traitées avec les catégories grammaticales en colonnes 3, 4 et 5 (séparées par un *point virgule* dans le champ NATURE de la base de données) nous permet de déduire les informations suivantes : 100 entrées avec des acceptions différentes, 1.832 entrées absentes du DES et 3.178 entrées mises à jour dans le DES.

Un point important déjà évoqué en 2.2 concernant la séparation des différentes catégories grammaticales d'une même entrée est à repréciser. Dans le fichier TLFi de départ, une même entrée apparaissait plusieurs fois complété d'un chiffre (1,2, ... 6) lorsqu'elle était considérée comme dégroupée. Elle figure donc dans l'interface publique du TLFi avec plusieurs onglets, et souvent avec plusieurs catégories grammaticales [10]. Dans le champ *nature* de la base de données du DES, les différentes catégories ont été juxtaposées et séparées par un point virgule sauf dans le cas de redondance. Par exemple *accusation* est considérée par le TLFi comme dégroupée (avec une origine étylomogique identique pour les deux), la première en tant qu'*action en justice*, la seconde, plus rare, comme *mise en évidence, accentuation*. Or dans les deux cas, il s'agit d'un substantif féminin. Le code "subst. fém." n'a donc été enregistré qu'une fois dans le DES.

Dans le cas où une entrée du fichier TLFi avait deux ou plusieurs colonnes avec des catégories grammaticales différentes, alors que cette entrée figure dans l'interface publique du DES avec un point d'entrée unique, ces catégories ont été ajoutées séparées également par un point virgule.

La notion d'entrée dégroupée du TLFi (telle qu'exprimée par un chiffre ajouté à la fin du mot) n'a pas été conservée quand il s'agit d'une même catégorie grammaticale (Voir 2.2). Il faut dire que cette notion diverge selon les dictionnaires. En effet, *accusation* n'est pas considérée avec deux origines étymologiques différentes selon le Grand Robert : une seule entrée en tant que substantif féminin se présente contrairement au TLFi.

Une autre remarque également : **le programme d'insertion en python a à ce stade été modifié**. En effet, la procédure vérifiant l'entrée dans la base de données du DES a été corrigée. Non seulement la vérification doit se faire sur le champ *graphie*, celle qui est affichée lors d'une recherche mais également sur le champ *cnrtl* qui donne la forme générique du mot (généralement au masculin). La correction de cette erreur a permis de mettre à jour 35 substantifs supplémentaires

[10] Par exemple https://www.cnrtl.fr/lexicographie/pointer

dans le DES mais n'a pas eu d'incidence sur les verbes et les adjectifs précédemment traités.

Après traitement, un total de 49.104 (44.094 + 5.010) sur 54.280 entrées du TLFi soit 90,4 % ont été utilisées, et dans le DES, 31.192 entrées avec le champ *nature* renseigné sur 50.451 soit 62%

2.8 Introduire une troisième catégorie de mots mélangés

Cette troisième catégorie de mots mélangés comme les précédentes récupère les entrées dont la colonne 2 correspond à *ale, ande, ane, ate, aude, euse, iale, ienne, ière, ieuse, ile, ine, ite, ive, orse, ose, ote, otte, ouse, oute, ue, une, ure, use, ute.*

Nous retenons 919 entrées avec jusqu'à 4 catégories grammaticales différentes (colonne 3 à 6 dans le fichier à traiter). Le fichier de log a malheureusement été détruit par erreur sans avoir été sauvegardé sur le serveur git.

Après traitement, un total de 50.023 (49.104 + 919) sur 54.280 entrées du TLFi soit 92% ont été examinées, et dans le DES, 31.757 entrées avec le champ nature ont été renseignées sur 50.451 soit 63%

2.9 Introduire une quatrième catégorie de mots mélangés

On arrive dans cette phase du traitement à des cas très particuliers et il faut donc prendre les lignes une à une. Il reste dans le fichier TLFi de départ 4.206 lignes à étudier. Nous avons pris en compte le début du fichier jusqu'à la ligne 2.515.

Les cas de mots avec plusieurs orthographes ont été pris en compte, les mots invariants, les prépositions, les interjections, onomatopées, ...

Sur les 2.619 entrées traitées, 92 étaient des entrées dégroupées, 1.686 étaient absentes du DES et 993 ont été renseignées dans le DES.

Après traitement un total de 52.642 (50.023 + 2.619) sur 54.280 entrées du TLFi soit 97 % ont été utilisées, et dans le DES, 32.503 entrées avec le champ nature renseigné sur 50.451 soit 64,4%

2.10 Introduire une cinquième catégorie de mots mélangés

Dans le fichier TLFi de départ, de nombreuses lignes sont intraitables (#NOM?,élément préf. ou élém. formant pour les mots avec « - »,...). Les lignes avec des parenthèses à la fin du mot (par exemple : BRUNANTE (À LA), CATIMINI (EN), CONTRE-BIAIS (À), CONTREBORD (À),...) ont été modifiées manuellement pour positionner le contenu entre parenthèses sans ces dernières devant le mot. Nous obtenons ainsi un fichier de 1.551 lignes.

Sur les 1.551 entrées traitées, 44 étaient des acceptions différentes, 949 étaient absentes du DES et 602 ont été renseignées dans le DES.

Après traitement, un total de 54.193 (52.642 + 1.551) sur 54.280 entrées du TLFi soit 99,8 % ont été utilisées, et dans le DES, 32.988 entrées avec le champ nature renseigné sur 50.451 soit 65,4%

2.11 Bilan intermédiaire

A ce stade, **au 17 mars 2021**, tout le contenu du fichier TLFi de départ a été pris en compte. Les lignes non exploitables sont dans le fichier TLFi_complet_lemmes_NonTraitable_20210317b.csv sur le git. Certaines font référence à d'autres entrées (COMMANDATURE voir KOMMANDANTUR. , ALKALIN voir ALCALIN....) . Il reste quelques locutions qui auraient pu être traitées mais qui l'ont été par d'autres moyens expliqués ci-dessous.

Il reste 17.463 entrées dans le DES pour lesquelles aucune catégorie grammaticale n'a été trouvée. Nous allons donc passer par d'autres moyens exposés ci-dessous.

3 Traitement semi-automatique sur les verbes

Le tri alphabétique de l'extraction des entrées du DES sans catégorie grammaticale avec 17.463 lignes, permet de s'apercevoir que 263 lignes commencant par s' et 1.147 lignes commençant par se, s'avèrent, après vérification, être des verbes. Une fois ces 1.411 lignes traitées en tant que verbe, il reste 16.052 lignes à étudier.

4 Utilisation de la librairie Spacy avec Python

La librairie *fr_ dep_ news_ trf* [11] est un pipeline de transformateurs français qui contient un ensemble de composants : morphologiseur, analyseur syntaxique, régleur d'attributs, lemmatiseur,.... (camembert-base). L'entrainement a été réalisé sur des données provenant de trois sources :

- UD_FrenchSequoia qui est une conversion automatique du corpus français Sequoia (French Sequoia corpus).
- Le modèle camembert-base basé sur le modèle RoBERTa. Il a été entrainé sur le corpus OSCAR(Open Super-large Crawled Aggregated coRpus)
- Des fichiers additionnels : spaCy lookups data

La première source provient de l'INRIA. Elle contient 3,099 phrases françaises de Europarl (parlement européen), du magazine Est Republicain, du Wikipedia français et de l'agence européenne de médecine. Le manuel d'annotations est disponible en ligne.

Le composant qui nous intéresse est celui qui va associer une catégorie grammaticale aux mots restants. En linguistique, l'étiquetage morpho-syntaxique, aussi appelé étiquetage grammatical ou **POS tagging (part-of-speech tagging)** est le processus qui consiste à associer aux mots d'un texte les informations grammaticales correspondantes comme la partie du discours, le genre, le nombre, etc. à l'aide d'un outil informatique. [12]
Nous avons donc commencé par extraire de la base MySQL du DES les entrées qui n'avaient pas de catégories grammaticales. Après quelques tests d'insertion de la catégorie grammaticale avec la librairie Spacy et un programme python [13], il s'avère que les mots composés et séparés par un espace ne sont pas pris en compte. Dans le tableur récupéré, nous avons donc les entrées sans espace : cela donne 10.139 entrées (fichier graphiesTraiteesSpacy.csv) pour lesquelles Spacy nous donne une catégorie grammaticale.

4.1 Première extraction

Nous avons trié le fichier graphiesTraiteesSpacy.csv sur la catégorie grammaticale et pour le code *POS VERB* nous récupérons les entrées finissant par *-er* et *-ir*. Une vérification permet de supprimer les entrées qui ne sont pas des verbes : *décrottoir, débirentier, parmentier*. Nous obtenons 588 lignes.

4.2 Seconde extraction

Toujours à partir du fichier graphiesTraiteesSpacy.csv, nous récupérons les entrées finissant par *é*. Plusieurs d'entre elles sont étiquetées comme verbe (code *VERB* dans Spacy), nous les remplaçons par *part. passé* sauf *abécédé, vulturidé* qui sont des substantifs et *ollé-ollé* une interjection.
Puis les lignes concernant des noms propres (code *PROPN*) et des ponctuations (code *PUNCT* ont été corrigées manuellement. Enfin les lignes avec des substantifs et des adjectifs (codes *NOUN* et *ADJ*) ont également été vérifiées.

Au total nous avons 951 entrées à traiter.

4.3 Troisième extraction

Le programme InsertSpacyCat.py est modifié pour tenir compte des tirets et des apostrophes. Nous récupérons ainsi 328 verbes commençant par *s'*, 404 adverbes et 275 substantifs finissant par *-ment*. Un total de 1007 lignes à traiter.

4.4 Quatrième extraction

Le programme InsertSpacyCat.py est modifié pour tenir compte des espaces. Mais cette modification n'apporte pas d'amélioration apparente dans la détection des catégories grammaticales par

[11] https://spacy.io/models/fr
[12] Définition sur Wikipedia : https://fr.wikipedia.org/wiki/Étiquetage_$morpho-syntaxique$
[13] Ce programme, InsertSpacyCat.py, importe les librairies fr_dep_news_trf et spacy, et pour chacune des 10.139 entrées, regarde si elle possède une catégorie grammaticale, (champ pos_) et créé un fichier de sortie de type .csv avec l'entrée et le code POS

Spacy.
Nous récupérons toutes les entrées qui commencent par *à*. Nous obtenons 362 lignes.
De façon générale, pour toute expression idiomatique plus ou moins figée commençant par *à* sont considérées comme adjectif si elles figurent à droite d'un substantif (*un projet à bas coût*) ou comme adverbe à droite d'un verbe ou d'un participe (*poursuivre un projet à marche forcée* ; *évaluer un coût à la louche*). Depuis quelques décennies on emploie les codes *ADJECTIF* et *ADVERBE* comme des catégories fonctionnelles au-delà de leur définition morphologique classique.
Nous avons choisi de tout étiqueter en adverbe et celles présentées sur cette page wiktionary [14] comme locution adjectivale ont été corrigées.

4.5 Cinquième extraction

On récupère toutes les lignes NOUN, ce qui donne un fichier de 4.608 lignes.
Nous vérifions quelques points : Nous trouvons 200 entrées finissant par *er, ir* et *dre* étiquetées verbe. Nous trouvons 534 entrées finissant par *-eur, -tre, -ire* et *-oir* étiquetées substantifs (sauf *stupéfaire*).
Les autres lignes étiquetées NOUN par Spacy passent dans le code subst.

4.6 Second bilan intermédiaire

Nous avons dans la base de données du DES 41.901 entrées avec le champ nature renseigné sur 50.496 soit 83%. Il reste 8.593 entrées à traiter.

5 Traitement manuel

5.1 Sixième extraction

Il reste 8.488 entrées dans le DES non renseignées.

Il nous a fallu étudier un peu plus attentivement le fichier afin de repérer des schémas répétitifs.

Avec l'aide de Jacques François, les entrées commençant par :

- *au, aux, à : au bénéfice de, au moment où, aux environs, ...*
- *de, sur, avec : de marbre, de manière à, avec plaisir, de bon coeur, de façon à, sur le coup, ...*
- *pour, par, en, sans : pour du beurre, sans crier gare, en direction de, par principe, ...*
- *dans, dès, du , hors, peu , sous, tout : dans le but de, dès que possible, du moment que, hors de combat, peu à peu, sous la main, tout de suite, ...*

suivies d'un espace pouvaient être considérées comme des adverbes. Nous avons 944 entrées.

De même que les entrées commençant par un verbe suivi d'un espace comme :

- *donner, être, faire : donner sa parole, être en pétard, faire son chemin, ...*
- *jeter, jouer, laisser : jeter l'ancre, jouer de malchance, laisser pour compte, ...*
- *lever, mettre, passer : lever l'ancre, mettre à la porte, passer à l'acte, ...*
- *porter, prendre, réduire : porter asssistance, prendre à coeur, réduire en poudre, ...*
- *remettre, rendre, reprendre : remettre à flot, rendre service, reprendre ses esprits, ...*
- *s'en, s' ou se : s'en sortir, se tenir à carreau, ...*
- *sortir, tenir, tirer : sortir de ses gonds, tenir au courant, tirer profit, ...*
- *tomber, tourner : tomber d'accord, tourner en ridicule, ...*

[14] Voir https://fr.wiktionary.org voir Locutions Adjectivales en français

sont étiquetés comme verbes.

Enfin, les mots commençant par :

- *homme : homme de loi,...*
- *jeu : jeu d'esprit,...*
- *lever : lever du jour, lever du soleil,...*
- *maison, maître, mauvais, mise : maison de correction, maître coq, mauvais oeil, mise au point*

sont étiquetés comme substantifs.

Nous obtenons en tout 2.197 lignes à traiter.

Après traitement, il reste 6.292 lignes à compléter.

5.2 Septième extraction

Les entrées commençant par :

- *aller : aller de pair, aller mieux,...*
- *avoir : avoir confiance, avoir envie,...*
- *battre , casser, changer, chercher, couper : battre la mesure, casser du sucre, changer d'air, chercher querelle, couper l'herbe sous le pied,...*
- *demander, devenir, dire, fermer, ficher, gagner,: demander grâce, dire ses quatre vérités, fermer sa gueule, ficher le camp, gagner du temps,...*
- *monter, montrer, ouvrir, payer, perdre, prêter : monter à bord, montrer son nez, ouvrir son coeur, payer de sa personne, perdre connaissance, prêter main-forte,...*

sont étiquetées comme verbes.

De même que :

- *art, corps, coup, champ : art culinaire, corps social, coup de filet, coup de foudre,champ de bataille, ...*
- *faux, femme, feu, fille, fils : fausse couche, femme de compagnie, feu follet,...*
- *fièvre, fleur, flux, force : fièvre jaune, fleur de l'âge, force navale,...*
- *gens, garde, grand/grande, gros : garde du corps, grande perche, gros mot,...*
- *herbe, jeune, jour, langue : jeune âge, jour de jeûne,...*
- *lettre, lieu, linge, liste, livre, loup : lettre de change, lieu sûr,...*
- *machine, mal, maladie, marchand, ministre : machine infernale, mal de poitrine, marchand de soupe, ministre du culte, ...*
- *pain, peau, petit, pierre, pièce, planche, poids, point, pot : pain de sel, petit doigt, pierre précieuse, pièce d'artillerie, planche à pain, point du jour, pot à beurre,...*
- *rat, sac, sens : sac de noeuds, sens moral,...*
- *terre, tour, tête, vieille, vieux : terre cuite, tour de scrutin,...*
- divers substantifs connus : *Babel, Christ,...*

sont étiquetés comme substantifs.

Celles commençant par *comme (comme il faut, comme quatre,...)* sont étiquetés *locution.*

1.236 lignes sont traitées.

5.3 Huitième extraction

L'idée dans cette huitième extraction est d'extraire les entrées qui commencent par le même mot : *agent (de change, de liaison,...), avant (l'heure, toute chose,...), bien (à propos, entendu,...), fil (à la patte, conducteur,...), mal (à l'aise, fichu,...), ...* On obtient ainsi 1.235 lignes.
Puis dans une seconde étape, nous affectons automatiquement à ces mots composés la catégorie grammaticale du premier mot si elle est déjà renseignée et nous revérifions manuellement le tout.

Une fois ces lignes traitées, il nous reste 3.818 entrées.

5.4 Neuvième extraction

Nous nous intéressons aux entrées finissant par *-ant* et *-iste*. Avec ces 558 nouvelles entrées ainsi détectées, nous introduisons les notions de **nom composé** et de **nom propre composé** qui apportent des précisions supplémentaires [15].

5.5 Les dix extractions suivantes

Avec les 3.260 entrées restantes, nous partons sur des traitements semi-automatiques :
Dans le tableur, nous récupérons 339 entrées avec les mots finissant par *-ique, -eur, -euse*.
Nous récupérons 635 entrées avec des mots contenant des espaces et dans le tableur, nous créons des colonnes pour les catégories grammaticales les plus courants et insérons le chiffre 1 dans la colonne correspondante. Une macro permet de transformer le chiffre 1 par le code grammatical retenu. Les diffentes catégories grammaticales à laquelle appartient l'entrée en question sont ensuite concaténées séparées par une virgule.
L'existence de l'entrée est vérifiée dans le TLFi et également dans le Grand Robert mis à disposition par le Service de Documentation (SCD) de l'Université de Caen. Cela a permis de supprimer certains mots non présents comme *broco, bixa, biclo, buxus*.
L'aide d'un stagiaire [16] a permis de traiter plus rapidement ces dernières entrées.

6 Vérifications

6.1 La cohérence des liaisons synonymiques par rapport aux catégories grammaticales

A l'aide d'un programme Python, nous lisons la base de données Mysql des synonymes. Pour *mot1* synonyme de *mot2*, nous avons appliqué les conditions présentées dans la table 8. Si la ligne traitée satisfait une de ces conditions, elle est mémorisée pour être étudiée manuellement.

mot1 est un verbe mais n'est pas un substantif, ni un adjectif, ni un adverbe, ni une locution et mot2 n'est ni un verbe, ni un adverbe, ni une locution
même condition que précédemment en échangeant mot1 et mot2
mot1 est un adverbe et un verbe
mot2 est un adverbe et un verbe
mot1 est un adverbe mais n'est pas un verbe et mot2 est un verbe mais n'est pas un adverbe

Table 8: Conditions finales programmées pour vérifier les relations synonymiques

Le programme a été exécuté une vingtaine de fois en améliorant les conditions de vérifications données dans la table 8. Cela a permis de corriger plus de 250 liaisons synonymiques.

Par exemple, *fier* était enregistré comme verbe uniquement alors qu'il est synonyme de *fort, hautain, noble,...* en tant qu'adjectif. De même pour *conseiller* enregistré comme verbe uniquement et non comme substantif. La liaison entre *dire du mal* et *ragot* a été supprimée. *Se souvenir*, noté comme substantif masculin, a été transformé en verbe. *Pourparler*, marqué comme verbe a été modifié en tant substantif masculin.

[15] Voir les différents types de substantifs sur ce site : https://www.cosmovisions.com/nom.htm
[16] Louis-Geoffroy Gousset, master 1, promotion 2021-22 en Science Du Langage

6.2 L'homogénéité des codes grammaticaux

Le champ *Nature* de certaines entrées avec un point virgule ont été vérifiées de façon à ce que les codes grammaticaux de part et d'autre du point virgule soient listés dans le même ordre. Par exemple, certaines entrées avaient *adj. et subst.;adj.* et d'autres *adj.;adj. et subst.* Toutes les permutations possibles jusqu'à quatre points virgules donc quatre acceptions différentes ont été étudiées. Un résultat de 54 paires du champ *Nature* a été détecté par programmation Python comme devant être vérifiées [17]. Cela donne 430 combinaisons de catégories grammaticales différentes comme l'indique le fichier catgram_20221108.csv [18] où apparaît pour chaque valeur du champ *Nature* le nombre d'entrées concernées.

6.3 Le traitement des entrées dégroupées

Dans l'insertion des mots mélangés dans les paragraphes 2.6, 2.7, 2.8, 2.9 et 2.10, le choix a été pris de séparer les codes grammaticaux dans les colonnes 3, 4 et 5 pour 2.6, 2.7 et les colonnes 3 à 6 pour 2.8. Or ces entrées avec leur différents codes gramaticaux **mais sur une ligne uniquement** ne sont pas considérées comme des entrées dégroupées par le TLFi (qui rappelons-le les représente sur plusieurs lignes en concaténant un chiffre de 1 à 9 à la fin du lemme). La liste de ces 844 mots est donnée dans le git [19]

6.4 Résultat final : quelques exemples

La table 9 nous montre quelques entrées du DES avec le champ *nature* renseigné.

Entrée	Nature
abattre	verbe trans.
accusation	subst. fém.
commettre	verbe;verbe trans. dir. et indir.;verbe trans.
meuble	subst. masc.;adj. et subst.;adj.
pique	subst. masc.;subst. fém.
faire	subst. masc.;verbe substitut.;verbe auxil.;verbe trans.
mise	part. passé et adj.;subst. fém.

Table 9: Extrait du résultat dans la base de données du DES

pique a 5 acceptions dans le TLFi en ligne et 3 dans le Grand Robert. Il est enregistré avec 2 catégories grammaticales recouvrant l'ensemble des acceptions. Toutes les précisions grammaticales données dans le fichier TLFi sont conservées comme nous pouvons le constater pour *commettre*.

7 Conclusion

Ce long travail a demandé de la rigueur dans les différentes étapes. Une partie du travail est disponible sur un dépot public https://git.unicaen.fr/crisco-des-public/descatgram. Le détail complet des opérations chronologiques ainsi que l'ensemble des fichiers et programmes nécessaires à l'élaboration de ce projet est présent sur un git privé pour l'instant. A ce stade, les catéories grammaticales insérées dans le DES n'apparaissent pas dans l'interface publique. Ce sera l'objet d'un autre projet.

[17] Voir https://git.unicaen.fr/crisco-des-public/descatgram/-/blob/master/VerifPermutationsCatGram_20221018.csv
[18] Voir https://git.unicaen.fr/crisco-des-public/descatgram/-/blob/master/catgram_20221108.csv
[19] Voir https://git.unicaen.fr/crisco-des-public/descatgram/-/blob/master/CatGramErreursAcceptions.csv