# 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

19 février 2024

**Résumé**

Ce document détaille la procédure suivie pour introduire la catégorie grammaticale dans la base de données du Dictionnaire Électronique des Synonymes (DÉS) [1] du laboratoire CRISCO [2] (50 457 entrées au 30 janvier 2024). **Ce travail a eu lieu en deux étapes. Une première de janvier 2021 à novembre 2022. Une seconde de juin 2023 à novembre 2023**.

La première étape s'est déroulée en deux grandes phases :
- à partir d'un premier jeu de 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*), transmis sous forme d'un classeur (au format .xlsx) contenant pour chaque entrée, les catégories grammaticales correspondantes. Avec des programmes en langage Python, les verbes, substantifs, adjectifs et adverbes ont été introduits dans la base de données du DÉS. Ensuite une recherche par schémas (*-ais, -euse, -ale, -ande,...*) avec plusieurs groupes de mots mélangés, suivie d'un traitement automatique a 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].

La seconde étape s'est déroulée en trois grandes phases :
- des corrections restantes à réaliser suite à la première étape,
- l'étude d'un second et ancien jeu de données de l'ATILF contenant 49 855 entrées avec les catégories grammaticales,
- l'étude d'un fichier plus récent (septembre 2023) de l'ATILF de 103 329 lignes.

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

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

# Table des matières

# 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 DÉS, en particulier, suite aux retours de plusieurs internautes fidèles [5].

Nous sommes donc partis de trois sources de données provenant de l'ATILF [6]. La première a été utilisée dans l'étape 1 et les deux suivantes dans l'étape 2.

Nous décrivons dans ce rapport les étapes techniques et les décisions prises pour traiter et récupérer de façon optimale les informations issues de ces trois sources.

# 2 Première étape de janvier 2022 à novembre 2022

## 2.1 A partir du fichier des lemmes du TLFi

### 2.1.1 Etude préliminaire du fichier

Ce fichier intitulé *TLFI complet lemmes.xls*, transmis à un chercheur du CRISCO lors d'une collaboration remontant à plusieurs années, a été le point de départ de ce projet. Il se présente sous la forme d'un classeur avec 54 280 lignes (entrées) lisible avec un tableur dont un extrait est présenté dans la table 1. Un premier examen manuel du premier onglet intitulé "TLFI complet" 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 l'exemple donné dans la table 2)

Le contenu de ce fichier correspond aux données affichées sur la page web du CNRTL (*Centre National de Ressources Textuelles et Lexicales*) [7] .

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

6. Mathieu Constant directeur actuel de l'ATILF a donné son accord en juin 2023 pour l'utilisation de ces fichiers par le CRISCO pour ses travaux de recherche.

7. A partir de la page d'accueil https ://www.cnrtl.fr puis, dans le menu, les onglets portail lexical puis lexicographie. L'accès direct est https://www.cnrtl.fr/definition/

| 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.

### 2.1.2 Introduire les verbes

Le fichier *TLFI complet lemmes.xls* contient un onglet *Verbes* que nous avons utilisé et sauvegardé dans un autre fichier tableur au format CSV [8]. Ce fichier intitulé *TLFI complet lemmes-verbes.csv* contient 6981 entrées de type verbe [9].

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 DÉS. 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*, 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 du DÉS. Il

8. Comma Separated Values, format recommandé pour l'interopérabilité des données
9. Ce nombre est légèrement supérieur à celui donné au paragraphe 2.1.1 "Etude préliminaire" donnant 6977 verbes, ce dernier ne tenant compte que de la colonne 2. Or la catégorie grammaticale peut être stockée dans les colonnes 3 à 7

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 DÉS.

| |
|---|
| 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 du DÉS, 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 DÉS.**

Au final 6 981 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 DÉS. Le détail des opérations concernant les verbes absents et ceux avec plusieurs acceptions est listé dans le fichier de log du git public du CRISCO [10].

### 2.1.3 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.* et les colonnes 3 et 4 vides. Nous obtenons ainsi le fichier *TLFI complet lemmes-adj.csv* avec 4 641 lignes [11].

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 DÉS est de 2 491.

On peut s'étonner que quasiment la moitié des adjectifs du fichier TLFi ne soit pas présente dans le DÉS. Si on regarde rapidement le fichier de log [12], on s'aperçoit que 637 d'entre eux se

10. https://git.unicaen.fr/crisco-des-public/descatgram/-/blob/master/CatGramLogInsertionVerbes.csv
11. Dans l'étue préliminaire, paragraphe 2.1.1, nous avons mentionné 5181 adjectifs et non 4641. La différence est due à des entrées qui ont d'autres catégories grammaticales dans les colonnes 3 et 4, non prises en compte ici. Elles sont traitées dans les paragraphes suivants.
12. Le fichier de log d'insertion des adjectifs est disponible sur le git public du DÉS : https://git.unicaen.fr/crisco-des-public/descatgram/-/blob/master/CatGramLogInsertAdj.csv

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 DÉS. 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 DÉS à l'avenir.

Un premier bilan intermédiaire nous permet de conclure que 6 982 verbes et 4 641 adjectifs du fichier TLFi ont été traités, soit un total de 11 623 sur 54 280 entrées, environ 21%. Dans le DÉS, 5 507 verbes et 2 491 adjectifs ont été mis à jour sur 50 350, soit 16%.

### 2.1.4 Introduire les substantifs

Comme pour le traitement des adjectifs, nous avons recréé un fichier *TLFI complet lemmes-ATraiter-20210107.csv* en supprimant les adjectifs (*TLFI complet lemmes-adj.csv* précédent). Nous obtenons 42 655 lignes. Seules les entrées avec la colonne 2 contenant "subst" et les colonnes 3 et 4 vides ont été retenues. Le traitement de ces 28 588 entrées restantes 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 DÉS et 19 143 ont été mis à jour dans le DÉS (Voir le fichier de log des substantifs absents ou ayant plusieurs acceptions [13]).

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

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

### 2.1.5 Introduire les adverbes

Nous exécutons la même procédure que précédemment, à savoir, la recréation d'un fichier *TLFI complet lemmes-ATraiter-20230712.csv* à partir du précédent en enlevant toutes les lignes traitées. Nous obtenons un fichier de 14 022 entrées. Nous extrayons de ce fichier les entrées contenant le code "adv." en colonne 2 et aucun code en colonnes 3 et 4. 943 adverbes ont été retenus et traités automatiquement par programme [14].

Nous avons pris en compte jusqu'à présent un ensemble 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 DÉS, 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%.

### 2.1.6 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 DÉS, seront séparées par un *point virgule* [15].

Sur les 2 940 entrées traitées, 62 étaient des acceptions différentes, 852 étaient absentes du DÉS et 2 088 ont été renseignées dans le DÉS [16].

Après traitement un total de 44 094 (41 154 + 2 940) sur 54 280 entrées du TLFi soit 81,2 % a été utilisé, et dans le DÉS, 28.155 entrées avec le champ *nature* renseigné sur 50.451 soit 55,8%.

### 2.1.7 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 2.1.2 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 complétée 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 [17]. 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*

14. Voir fichier de log avec les adverbes absents du DÉS ou ayant plusieurs acceptions donc non pris en compte https://git.unicaen.fr/crisco-des-public/descatgram/-/blob/master/InsertAdvBD-Log.csv
15. Voir le paragraphe 2.1.2
16. Fichier de log : https://git.unicaen.fr/crisco-des-public/descatgram/-/blob/master/CatGramLogInsertMotsMelanges-1.csv
17. Par exemple https://www.cnrtl.fr/lexicographie/pointer

*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 DÉS.

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 DÉS avec un point d'entrée unique, ces catégories ont été ajoutées séparées également par un point virgule. Ce cas de figure a été revu au paragraphe 2.5.3.

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 2.1.2). 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 DÉS 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 dans le DÉS 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 % a été utilisé, et dans le DÉS, 31 192 entrées avec le champ *nature* renseigné sur 50 451 soit 62% [18]

### 2.1.8 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 DÉS, 31 757 entrées avec le champ nature ont été renseignées sur 50.451 soit 63%

### 2.1.9 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.

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

Les entrées de type "élém. formant" ont été ignorées (voir des exemples dans la table 8)

| | | | |
|---|---|---|---|
| NYCT- | NYCTI- | NYCTO- | élém. formant |
| OCT(A)-<br>OCULI- | OCTI-<br>OCULO- | OCTO-<br>élém. formant | élém. formant |
| HODO- | ODO- | élém. formant | |
| OENI- | OENO- | élém. formant | |

TABLE 8 – Exemples d'entrées de type "élem. formant" dans le fichier TLFi

Sur les 2 619 entrées traitées, 92 étaient des entrées dégroupées, 1 686 étaient absentes du DÉS et 993 ont été renseignées dans le DÉS.

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 DÉS, 32 503 entrées avec le champ nature renseigné sur 50 451 soit 64,4% [19]

18. Fichier de log : https://git.unicaen.fr/crisco-des-public/descatgram/-/blob/master/CatGramLogInsertMotsMelanges-2.csv
19. Fichier de log : https://git.unicaen.fr/crisco-des-public/descatgram/-/blob/master/CatGramLogInsertMotsMelanges-4.csv

### 2.1.10 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 DÉS et 602 ont été renseignées dans le DÉS.

Après traitement, un total de 54 193 (52 642 + 1 551) sur 54 280 entrées du TLFi soit 99,8 % a été utilisé, et dans le DÉS, 32 988 entrées avec le champ "nature" renseigné sur 50 451 soit 65,4% [20].

### 2.1.11 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 DÉS pour lesquelles aucune catégorie grammaticale n'a été trouvée. Nous avons donc étudié d'autres moyens exposés ci-dessous.

## 2.2 Traitement semi-automatique sur les verbes

Le tri alphabétique de l'extraction des entrées du DÉS 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.

## 2.3 Utilisation de la librairie Spacy avec Python

La librairie *fr_dep_news_trf*[21] est un pipeline de transformateurs français qui contient un ensemble de composants : morphologiseur, analyseur syntaxique, régleur d'attributs, lemmatiseur,....

L'entrainement a été réalisé sur des données provenant de trois sources :

— UD_FrenchSequoia [22] qui est une conversion automatique du corpus français Sequoia (French Sequoia corpus).

— Le modèle camembert-base [23] 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 footnotehttps://github.com/explosion/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 [24].

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. [25]

Nous avons donc commencé par extraire de la base MySQL du DÉS 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 [26], il s'avère que les mots composés et séparés par un es-

20. Fichier de log : https://git.unicaen.fr/crisco-des-public/descatgram/-/blob/master/CatGramLogInsertMotsMelanges-5.csv
21. https://spacy.io/models/fr
22. https://github.com/UniversalDependencies/UD_French-Sequoia
23. https://huggingface.co/almanach/camembert-base
24. https://gitlab.inria.fr/sequoia/deep-sequoia/-/blob/master/tags/sequoia-9.2/README-distrib.md
25. Définition sur Wikipedia : https://fr.wikipedia.org/wiki/Étiquetage_morpho-syntaxique
26. 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

pace 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 dans le git privé) pour lesquelles Spacy nous donne une catégorie grammaticale. De ce fichier, cinq types d'extractions ont été effectués (cf ci-dessous) avec pour chacun des vérifications manuelles.

### 2.3.1 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.

### 2.3.2 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.

### 2.3.3 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.

### 2.3.4 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 Spacy.
Nous récupérons toutes les entrées qui commencent par *à*. Nous obtenons 362 lignes.
De façon générale, toute expression idiomatique plus ou moins figée commençant par *à* est considérée comme adjectif si elle figure 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 *adj.* et *adv.* 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 [27] comme locution adjectivale ont été corrigées.

### 2.3.5 Cinquième extraction

On récupère toutes les lignes NOUN, ce qui donne un fichier de 4 608 lignes.

Nous vérifions la catégorie grammaticale de 200 entrées finissant par *er, ir* et *dre* : 32 sont en réalité des *verbes*. Nous vérifions 534 entrées finissant par *-eur, -tre, -ire* et *-oir* étiquetées substantifs : une est un verbe (*stupéfaire*).
Les autres lignes étiquetées NOUN par Spacy passent dans le code *subst.*

### 2.3.6 Second bilan intermédiaire

Nous avons dans la base de données du DÉS 41 901 entrées avec le champ "nature" renseigné sur 50 496 soit 83%.

27. Voir https://fr.wiktionary.org voir Locutions Adjectivales en français

## 2.4 Traitement manuel

### 2.4.1 Sixième extraction

Il reste 8 488 entrées dans le DÉS 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 ont pu 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, ...*

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.

### 2.4.2 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, devenir dingue, 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 étiquetées *locution*.

1 236 lignes sont traitées.

### 2.4.3 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.

### 2.4.4 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 [28].

### 2.4.5 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 également 635 entrées avec des mots contenant des espaces et dans le tableur, nous créons des colonnes pour les catégories grammaticales les plus courantes et insérons le chiffre 1 dans la colonne correspondante. Une macro permet de transformer le chiffre 1 par le code grammatical retenu. Les différentes 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 [29] a permis de traiter plus rapidement ces dernières entrées.

## 2.5 Vérifications

### 2.5.1 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 9. Si la ligne traitée satisfait une de ces conditions, elle est mémorisée pour être étudiée manuellement.

Le programme a été exécuté une vingtaine de fois en améliorant les conditions de vérifications données dans la table 9. 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

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

comme substantif masculin, a été transformé en verbe. *Pourparler*, marqué comme verbe a été modifié en tant substantif masculin.

| |
|---|
| mot1 est un verbe mais n'est pas un substantif, ni un adjectif, ni un adverbe, ni une locution<br>et<br>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<br>et<br>mot2 est un verbe mais n'est pas un adverbe |

TABLE 9 – Conditions finales programmées pour vérifier les relations synonymiques

### 2.5.2 L'homogénéité des codes grammaticaux

Le champ "nature" de certaines entrées avec un point virgule a été vérifié 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 [30]. Cela donne 430 combinaisons de catégories grammaticales différentes comme l'indique le fichier catgram_20221108.csv [31] où apparaît pour chaque valeur du champ *Nature* le nombre d'entrées concernées.

### 2.5.3 Le traitement des entrées dégroupées

Dans l'insertion des mots mélangés dans les paragraphes 2.1.6, 2.1.7, 2.1.8, 2.1.9 et 2.1.10, le choix a été fait de séparer par des points virgules les codes grammaticaux présents dans plusieurs colonnes d'une même ligne (entrée) du fichier TLFi. Cela revenait à considérer qu'il y avait plusieurs acceptions. Or ces entrées avec leur différents codes grammaticaux **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 avec les catégories grammaticales de l'ATILF est donnée dans le git [32]. **Ces entrées devront donc être réétudiées pour corriger cette erreur**.

### 2.5.4 Résultat final : quelques exemples

La table 10 nous montre quelques entrées du DÉS 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 10 – Extrait du résultat dans la base de données du DÉS

*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*.

30. Voir https://git.unicaen.fr/crisco-des-public/descatgram/-/blob/master/VerifPermutationsCatGram_20221018.csv
31. Voir https://git.unicaen.fr/crisco-des-public/descatgram/-/blob/master/catgram_20221108.csv
32. Voir https://git.unicaen.fr/crisco-des-public/descatgram/-/blob/master/CatGramErreursAcceptions.csv

# 3 Seconde étape de juin 2023 à novembre 2023

Cette seconde étape a eu lieu de juin à novembre 2023. Tout d'abord (phase 1), nous avons repris les 844 entrées restantes à corriger (cf 2.5.3 ). Puis nous avons complété notre base de données avec deux autres fichiers provenant de l'ATILF [33], le premier assez ancien échangé lors d'une collabotation CRISCO-ATILF datant de plusieurs années (phase 2), l'autre envoyé par l'ATILF en septembre 2023 (phase 3).

## 3.1 Corrections restantes à réaliser suite à la première étape (phase 1)

Pour traiter ces 844 entrées présentes dans le fichier CatGramErreursAcceptions.csv [34], nous avons réalisé un programme python qui créé un fichier résultat avec les champs suivants :

- une première colonne contenant l'entrée du fichier traité
- une seconde colonne qui contient :
  - "non présente dans le DÉS : " si c'est le cas ou
  - "catgram identique" suivi de la catégorie grammaticale dans le DÉS ou
  - "cat gram identiques mais ordre différent (catgram DÉS avec ; catgram TLFI avec , : " suivi de la catégorie grammaticale dans le DÉS ou
  - la catégorie grammaticale dans le DÉS uniquement si les 3 conditions précédentes ne sont pas vérifiées
- une troisième colonne avec la catégorie grammaticale dans le fichier TLFi traité.

Le fichier résultat est sur le git public du DÉS [35].

La table 11 nous en présente un extrait.

| colonne 1 | colonne 2 | colonne 3 |
|---|---|---|
| abstrait | cat gram identiques mais ordre différent (catgramDES avec ; catgramTLFI avec , ) : adj. et subst. masc. ;part. passé | part. passé,adj. et subst. masc. |
| bénéficiant | non présente dans le DES | |
| droit | adj. et subst. ;adv. et subst. ;adj. ;subst. masc. | adj.,adv. et subst. |
| engluage | catgram identique :subst. masc. | subst. masc. |
| extraterrestre | catgram identique :adj. et subst. | adj. et subst. |
| tonifiant | cat gram identiques mais ordre différent (catgramDES avec ; catgramTLFI avec , ) : adj. et subst. masc. ;part. prés. | part. prés.,adj. et subst. masc. |

TABLE 11 – Extrait du fichier résultat de la 2nde étape - phase 1

Nous avons :

- Cas 1 : 208 entrées non présentes dans le DÉS
- Cas 2 : 65 entrées avec des catégories grammaticales différentes
- Cas 3 : 548 entrées avec des combinaisons de catégories grammaticales identiques mais non enregistrées dans le bon ordre : avec des « ; » comme séparateur coté DÉS ( incorrect) et coté TLFi avec « , » comme séparateur ( correct)
- Cas 4 : 23 catégories grammaticales uniques identiques dans le TLFi et dans le DÉS

Ce qui nous donne bien un total de 844 entrées.

Les 548 entrées du cas 3 peuvent être traitées automatiquement. Les 23 du cas 4 n'ont pas besoin d'être traitées car elles sont correctes dans le DÉS. Les 65 du cas 2 doivent être traitées manuellement.

Un nouveau fichier de type tableur est créé avec les 613 entrées ( cas2 + cas3) et leur nouvelle catégorie grammaticale. Un programme ensuite lit ce fichier et remplace l'ancienne valeur par la nouvelle dans la base de données du DÉS.

33. https://atilf.fr
34. Voir note précédente
35. https://git.unicaen.fr/crisco-des-public/descatgram/-/blob/master/CatGramErreursAcceptionsRecup_2023-07-13.csv

## 3.2 Premier fichier de l'ATILF (phase 2)

Il s'agit plus exactement de 81 fichiers au format XML. Dans le cadre du projet de modélisation graphique des notices historiques du TLFi [36] [37] [38] [39], un programme a été créé pour extraire les données XML et les enregistrer au format excel (xlsx).

Pour notre projet, nous sommes donc partis de ces fichiers excel pour en créer un unique avec l'entrée et sa catégorie grammaticale TLFi.

Une extraction de la base de données du DÉS est réalisée dans un fichier au format tableur csv avec deux colonnes : l'entrée et sa catégorie grammaticale DÉS.

Un programme en python pour comparer les deux fichiers est créé. Il donne le résultat suivant :

— Nombre entrées dans le DÉS 50 420
— Nombre entrées dans le TLFi 49 854
— Nombre d'entrées en commun DÉS- TLFi : 24 210
— Nombre d'entrées en commun avec la même catégorie grammaticale (code 1) : 23 548
— Nombre d'entrées en commun avec les cat gram du DÉS incluses dans TLFI (code 2) : 449
— Nombre d'entrées en commun avec les cat gram du DÉS différentes du TLFI (code 3) : 213
— Nombre d'entrées dans le DÉS absentes du TLFi (code 4) : 26 209
— Nombre d'entrées dans le TLFi absentes du DÉS (code 5) : 25 644

Sur les 449 entrées référencées dans le cas de figure "Nombre d'entrées en commun avec les cat gram du DÉS incluses dans TLFI (code 2)", seules 91 sont à traiter manuellement, les autres étant dues à l'ajout d'un point à la fin de la catégorie verbe dans le TLFi.

Les entrées modifiées manuellement dans le DÉS sont :

— *amoncellement, café, sésame , carabinier , maroquinier, persiflage* : "subst." est remplacé par "subst. masc."
— *plan-plan* : "adv." est remplacé par "loc. adj. et adv."
— *moulinette, hamada, datation, rétrospection, ursuline, nécropsie, ambiguïté, endémie* : "subst." est remplacé par "subst. fém."
— *université, trinité, gueuse, asse, cité* : "subst. fém. ;subst." est remplacé par "subst. fém."

Les autres cas concernent des différences entre "," et " ;". Une vérification dans le TLFi permet de savoir s'il s'agit d'une acception ou pas.

Concernant les 213 entrées référencées dans le cas de figure "Nombre d'entrées en commun avec les cat gram du DÉS différentes du TLFI (code 3)", 93 d'entre elles sont une inversion, par exemple "adj. et subst." d'un coté et "subst. et adj." de l'autre.

Les 120 restantes ont été vérifiées manuellement en expliquant les choix faits par une colonne supplémentaire dans un nouveau fichier de comparaison [40]. Il y a, par exemple, des espaces manquants dans le TLFi (pour *mobile*), des différences d'abbréviations (invar. / inv. pour *faire-part*).

## 3.3 Second fichier de l'ATILF de septembre 2023 (phase 3)

Ce fichier plus récent de l'ATILF contient 103.328 lignes (il y en avait 49.854 dans le précédent fichier). Il est constitué de six colonnes intitulées articleID, parentID, source, content, category, gender, feminine. La table 12 donne quelques exemples d'entrées.

On remarque que certaines entrées (colonne content) sont présentes sur plusieurs lignes, probablement liées à la notion d'acception [41].

Un programme de comparaison entre une extraction de la base de données du DÉS et ce fichier donne les résultats suivants :

— Nombre entrées dans le DÉS 50 434

36. https://crisco.unicaen.fr/recherche/thematique-2-recherche-linguistique-appliquee-a-des-pratiques-et-a-la-production-de-ressources-electroniques/
37. https://hal.science/hal-04052295
38. https://hal.science/hal-04133117
39. https://hal.science/hal-04052311
40. https://git.unicaen.fr/crisco-des-public/descatgram/-/blob/master/compar_CatGram_2023-08-23_code3AVerifier.csv
41. La notion d'acceptions est gérée différemmant suivant les dictionnaires. Par exemple, pour *accusé*, le Grand Robert n'a qu'une page en tant que nom et adjectif alors que le TLFi en a deux : Voir https://www.cnrtl.fr/lexicographie/accusé

| articleID | parentID | source | content | category | gender | feminine |
|---|---|---|---|---|---|---|
| 87 | 87 | source,parsed | abaissant | adjectif | | abaissante |
| 971 | 971 | source,parsed | accusé | nom | | accusée |
| 972 | 972 | source,parsed | accusé | adjectif | | accusée |
| 972 | 972 | source,parsed | accusé | nom | | accusée |
| 998 | 974 | source | grimace | nom | féminin | |

TABLE 12 – Extrait du fichier TLFi de septembre 2023

— Nombre entrées dans le TLFi2 103 328
— Nombre d'entrées unique dans le TLFi2 : 89 392
— Nombre d'entrées en commun DÉS- TLFi2 : 37 427
— Nombre d'entrées dans le DÉS absentes du TLFi2 : 13 007
— Nombre d'entrées dans le TLFi2 absentes du DÉS : 51 965

Comme exemples d'entrées dans le DÉS absentes du TLFi2, nous avons des expressions comme *sale type, en mauvais état, en matière de, etc...*, des verbes à la forme pronominale comme *s'élancer, s'introduire, etc...*

Présents dans le TFLi2 mais absentes du DÉS, nous avons principalement des suffixes comme *-triche, -variant, -sion*, des mots spécifiques sans synonynes comme *amphiprion, anisopode,etc ...*, des mots qui pourraient y être comme *pondérément*, enfin des éléments de composition comme *porte-* et tous ceux qui commencent par cet élément comme *porte-greffe, porte-malle, porte-montre, etc ....*

Le résultat du programme de comparaison créé un fichier tableur avec les colonnes suivantes :
— Entrée
— Catégorie grammaticale dans le DÉS
— Catégorie grammaticale dans le TLFi
— Une colonne intitulée "ok ?" avec la valeur True ou False suivant si les 2 catégories se *ressemblent* ou pas pour les 37 427 entrées communes

Les codes pour définir les catégories grammaticales dans les 2 fichiers ne sont pas identiques. Nous avons donc défini les règles suivantes pour traiter cette *ressemblance* :

```
Si "verbe" est présent dans le DÉS et le TLFi
ou
Si "subst." est présent dans le DÉS et "nom" dans le TLFi
ou
si "adj. est présent dans le DÉS et "adjectif" dans le TLFi
ou
si "adv." est présent dans le DÉS et "adverbe" dans le TLFi
ou
si "loc." est présent dans le DÉS et "locution" dans le TLFi
ou
si "interj." est présent dans le DÉS et "interjection" dans le TLFi
ou
si "prép" est présent dans le DÉS et "préposition" dans le TLFi

alors la colonne "ok?" est égale à True

sinon la colonne "ok?" est à False
```

Nous obtenons ainsi 336 lignes à vérifier manuellement c'est-à-dire avec la colonne "ok ?" à False. Ces lignes sont copiées dans un autre fichier tableur en y ajoutant deux autres colonnes : "corrections dans le DES" et "remarques pour l'ATIF".

237 d'entre elles ont du être modifiées et plusieurs ont été reportées auprès de l'ATILF comme étant à corriger de leur côté. La table 13 en donne un extrait.

On remarque un nombre important d'entrées marquées uniquement comme participe passé alors qu'elles sont aussi adjectif.

## 3.4 Vérification

Dans cette étape, nous nous sommes inspirés du programme du paragraphe 2.5 et les règles suivies de la table 9 pour extraire les lignes à contrôler. Or, ces règles sont partielles d'une part et

| Entrée | Catégorie grammaticale dans le DÉS | Catégorie grammaticale dans le TLFi | ok ? | corrections dans le DES | remarques pour l'ATIF |
|---|---|---|---|---|---|
| aérospatial | subst. | adjectif | False | adj., subst. fém. | |
| de guingois | adv. | locution | False | loc. | |
| de plain-pied | adv. | locution | False | loc. | |
| ferté | part. passé | nom | False | subst. fém. | |
| inventé | part. passé | adjectif | False | part. passé, adj. | |
| médiatique | adj. | nom | False | | adj. |
| porte à porte | subst. fém. | adverbe | False | loc. | |
| rapproché | part. passé | nom | False | part. passé, adj. | part. passé, adj. |
| toussotement | adv. | nom | False | subst. | |

TABLE 13 – Extrait des corrections sur les 237 entrées répertoriées

inexactes d'autre part. Par exemple, une des règles vérifiait si le mot1 est un adverbe et le mot2 un verbe ou inversement. Si c'était le cas, il fallait vérifier la liaison synonymique. Or *super* est bien à la fois les deux avec en tant que verbe le sens de *aspirer, gober* et adverbe dans l'exemple *un endroit super / extrêmement calme.*

Nous sommes donc repartis sur une règle simple : pour deux mots synonymes, mot1 et mot2, si une des catégories grammaticales de l'un est présente dans l'autre, alors nous ne vérifions pas la liaison synonymique. Une première exécution sur cette base mais en ajoutant les exceptions de la table 14 donne 5 828 lignes à vérifier concernant 3 057 mots, ce qui est impossible à vérifier.

| |
|---|
| si mot1 et mot2 sont des verbes, des substantifs, des ajectifs, des adverbes, des locutions, des participes passé<br>ou<br>si mot1 est un nom et mot2 un substantif ou inversmeent<br>ou<br>si mot1 est adjectif et mot2 participe passé ou inversement<br>ou<br>mot1 est un adverbe et un verbe<br>ou<br>mot1 est adverbe et mot2 locution ou inversement |

TABLE 14 – 1er groupe d'exceptions appliquées pour la vérification

De nouvelles exceptions détaillées dans la table 15 ont donc été ajoutées.

| |
|---|
| si mot1 est un adjectif et mot2 un adverbe ou inversement<br>ou<br>si mot1 est adjectif et mot2 une locution ou inversement<br>ou<br>mot1 est un adjectif et mot2 un substantif<br>ou<br>mot1 et mot2 sont tous les deux des prépositions ou des interjections<br>ou<br>mot1 est un verbe et mot2 une locution<br>ou<br>mot1 est une locution et mot2 un substantif |

TABLE 15 – 2nd groupe d'exceptions appliquées ajoutées au 1er pour la vérification

En modifiant le programme pour tenir compte de toutes ces exceptions, nous obtenons un fichier de 725 lignes à vérifier. Dans ce fichier traité manuellement nous y ajoutons 3 colonnes : correction de la catégorie grammaticale du mot1, correction de la catégorie grammaticale du mot2 et autres corrections (par exemple suppression de la liaison synonymique).

Le fichier VerifLiaisonsSynoAntoAvecCatGram_2024-2-7-16-0.csv [42] contient les résultats.

42. https://git.unicaen.fr/crisco-des-public/descatgram/-/blob/master/VerifLiaisonsSynoAntoAvecCatGram_2024-2-7-16-0.csv

La table 16 donne quelques exemples corrigés.

| mot1 | cat gram mot1 | mot2 | cat gram mot2 | correction cat gram mot1 | correction cat gram mot2 | autres corrections |
|---|---|---|---|---|---|---|
| alternativement | adv. | coup sur coup | subst. masc. | | loc. adv. | |
| beaucoup | adv. | force | subst. fém. | | | liaison supprimée |
| clouer le bec | verbe | en boucher un coin | adv. | loc. verb. | loc. verb. | |
| en quarantaine | adv. | interdit | subst. masc. | | adj., subst. masc. part. passé | |
| faute de quoi | subst. fém. | sans quoi | adv. | loc. | | |
| frottée | subst. fém. | roulée | part. passé | | subst. fém. | |
| nature | subst. fém. | naturellement | adv. | | | liaison supprimée |
| sans arrêt | adv. | nuit et jour | subst. fém. | | loc. | |

TABLE 16 – Exemples d'entrées corrigées parmi les 725 détectées

# 4 Conclusion

Ce long travail a demandé de la rigueur dans les différentes étapes, même si nous sommes conscients qu'il reste des vérifications et des corrections à effectuer. Une partie du travail est disponible sur un dépot public [43]. 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 à l'université de Caen. A ce stade, les catégories grammaticales insérées dans le DÉS n'apparaissent pas dans l'interface publique mais elles sont présentes dans la dernière version déposée sur la plate-forme ORTOLANG [44].

43. https://git.unicaen.fr/crisco-des-public/descatgram
44. https://www.ortolang.fr/fr/accueil/