DEMC Journal

N° 2 | 2024
2024

Insertion des catégories grammaticales dans le Dictionnaire Électronique des Synonymes (DÉS)

Laurette Chardon
Ingénieure de recherche
UFR HSS
CRISCO
Université de Caen Normandie

Édition électronique :
URL :
https://demc-journal.org/articles/revue-2/3768-insertion-des-categories-grammaticales-dans-le-dictionnaire-electronique-des-synonymes-des
ISSN : 3036-5295
Date de publication : 15/02/2024

Pour **citer cette publication** : Chardon, L. (2024). Insertion des catégories grammaticales dans le Dictionnaire Électronique des Synonymes (DÉS). *DEMC Journal*, (2). https://demc-journal.org/articles/revue-2/3768-insertion-des-categories-grammaticales-dans-le-dictionnaire-electronique-des-synonymes-des

Cet article détaille la procédure suivie pour introduire la catégorie grammaticale dans la base de données du Dictionnaire Électronique des Synonymes (DÉS) du laboratoire CRISCO (50 457 entrées au 30 janvier 2024). Ce travail a eu lieu en deux étapes : une première de janvier à novembre 2022 et une seconde de juin à novembre 2023. La première étape de janvier à novembre 2022 s'est déroulée en deux grandes phases. A partir d'un premier jeu de données de l'ATILF (Analyse et Traitement Informatique de la Langue Française, UMR7118), laboratoire du CNRS 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. A partir d'autres moyens : le modèle d'intelligence artificielle (IA) fr_dep_news_trf disponible avec la librairie Spacy en langage Python va permettre au moyen de l'étiquetage morpho-syntaxique d'associer aux entrées du DÉS des informations grammaticales. La seconde étape de juin à novembre 2023 s'est déroulée en deux grandes phases : 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. Cet article s'appuie sur un document de travail, accompagné d'un dépôt git public.

This article details the procedure followed to introduce the grammatical category into the CRISCO laboratory's Dictionnaire Électronique des Synonymes (DÉS) database (50,457 entries, January 30th 2024). This work was carried out in two stages: the first from January to November 2022 and the second from June to November 2023. The first stage, from January to November 2022, took place in two main phases. First using an initial dataset from ATILF (Analyse et Traitement Informatique de la Langue Française, UMR7118), the CNRS laboratory responsible for maintaining and developing the TLFi (Trésor de la Langue Française informatisé), transmitted in the form of a workbook (.xlsx format) containing the corresponding grammatical categories for each entry. Using Python language programs, verbs, nouns, adjectives and adverbs were entered into the DÉS database. This was followed by a search using patterns (-ais, -euse, -ale, -ande, ...) with several groups of words mixed together. Then by other means: the artificial intelligence (AI) model fr_dep_news_trf available with the Spacy library in the Python language will use morpho-syntactic labelling to associate grammatical information with DES entries. The second stage, from June to November 2023, took place in two main phases: the study of a second, older ATILF dataset containing 49,855 entries with grammatical categories ; the study of a more recent ATILF file (September 2023) containing 103,329 lines. This article is based on a working document, accompanied by a public git repository.

Mots-clefs :
Dictionnaire, Synonyme, Catégorie grammaticale, DÉS, CRISCO, Dictionary, Synonym, Grammatical category

Préambule

Cet article fait référence à plusieurs fois à un document de travail (Chardon, 2024) et son dépôt git associé[1].

Introduction

Nous allons tout d'abord présenter l'historique et les principes de base du DÉS puis expliquer l'intérêt de ce projet.

Le Dictionnaire électronique des synonymes du CRISCO (DÉS) contient aujourd'hui plus de 50 000 entrées, 210 000 relations synonymiques et 32.000 liaisons antonymiques. La base de départ concernant les synonymes a été constituée à partir de sept dictionnaires classiques. Un premier travail, réalisé par l'INALF (Institut National de la Langue Française), a permis d'en extraire les relations synonymiques. Le laboratoire ELSAP, qui est devenu par la suite le CRISCO, a ensuite concaténé, homogénéisé et symétrisé les données. Depuis 1994, un important travail de correction se perpétue par l'ajout ou la suppression de liens synonymiques et antonymiques.

Après livraison des données de l'INALF, ce projet a démarré sous la responsabilité de Bernard VICTORRI, directeur de recherche, et Sabine PLOUX, ingénieure de recherche, qui ont défini et mis au point les principes de fonctionnement du dictionnaire : union des différentes ressources, symétrisation (générant plus de 50 % de relations supplémentaires), algorithme de calcul des cliques. La représentation spatiale de l'espace sémantique et sa projection sur un plan par calcul matriciel ont également été conçues et réalisées.

De nombreuses personnes sont intervenues, depuis, à la fois pour des corrections, la maintenance et des améliorations (voir la rubrique Historique sur la page de présentation du site web) parmi lesquelles Jean-Luc MANGUIN, Michel MOREL et Laurette CHARDON, ingénieurs successifs en charge du projet.

Les deux principes de base du DÉS est la symétrisation et la contextualisation : un mot A peut être remplacé par un mot B dans un contexte donné, sans modification

significative du sens. Comme exemples, nous avons :

- Les enfants jouent / s'amusent dans la cour.
- Marie joue/imite/simule Andromaque.
- Un vin âcre/âpre/vert.
- Des mots crus/verts

Suite à la mise en ligne publique, gratuite et sans publicité, dans les années 2000 du projet, le grand public s'est approprié l'interface d'affichage des synonymes (beaucoup de rédactrices et rédacteurs d'articles de blogs, de journaux, de livres ... Il suffit de lire les retours dans le livre d'or et les statistiques d'accès). L'espace sémantique[2] au cœur du projet scientifique était beaucoup moins utilisé (certainement en raison d'un manque de communication du CRISCO et à un manque de compréhension de la part des usagers, qu'ils soient issus du monde de la recherche ou pas).

Ceci dit, la recherche sur la base du DÉS s'est depuis étoffée avec des algorithmes de regroupement bien connus en théorie des graphes (Chardon, 2020).

Dans les années 2000, un autre projet intitulé "les atlas sémantiques"[3] développé par Sabine Ploux (co-fondatrice du projet) a vu le jour sur la base des données du DÉS et étendu à l'anglais, l'espagnol et le portugais.

L'ensemble des publications en relation avec le DÉS est regroupé dans une collection HAL[4] avec 71 entrées dont 56 publications avec le texte intégral. Les textes fondateurs sont ceux de Bernard Victorri, Sabine Ploux et Jean-Luc Manguin (Ploux, 1995, 1996 ; Ploux, Victorri, 1998 ; Victorri, Manguin, 1999). Les textes sur l'exploitation du DÉS depuis 2000 sont également dans la collection CRISCO-DES. Quant aux mises à jour mensuelles, elles sont sur la page de présentation du DÉS.

Compléter la base des synonymes du DÉS avec la catégorie grammaticale apporte des avantages très intéressants :

- Cela facilite la recherche de mots polysémiques. En effet ces derniers sont très souvent attachés à plusieurs catégories grammaticales. Obtenir une liste exhaustive de tels mots répond aux besoins des linguistes en particulier en diachronie (l'évolution du sens d'un mot dans le temps) ou pour l'étude d'homonymes qui n'ont aucune origine étymologique commune. La recherche et l'extraction de tels mots est ainsi simplifiée. Par exemple *mousse* est un adjectif (qui n'est pas tranchant : couteau à pointe *mousse*), un substantif féminin (la

mousse du gel-douche dans la baignoire) et substantif masculin (jeune garçon sur un navire)

- Cela permet également la vérification des liens synonymiques et antonymiques : soit la relation synonymique est une erreur (*en-tout-cas* adverbe était synonyme de *parapluie* substantif masculin par erreur) soit il manque une catégorie grammaticale à l'un des 2 mots (*fier* était enregistré comme verbe uniquement alors qu'il est synonyme de *fort, hautain, noble*. . . en tant qu'adjectif)

- Il est ainsi plus facile de repérer les acceptions. Dans le TLFi, un mot qui a plusieurs acceptions se traduit par plusieurs onglets car les sens sont disjoints (par exemple pour *canon*, *un tir de canon* n'a aucun rapport avec le *droit canon*). Lorsque les différents sens d'un mot découlent les uns des autres, il n'y aura qu'une seule entrée (le verbe *gagner* a une seule entrée avec différents sens : *acquérir quelque chose* mais aussi *mériter une récompense* ou encore *gagner le large, partir* ...). Il faut noter toutefois que les dictionnaires divergent sur la notion d'acception. Nous avons gardé dans notre base cette information telle qu'utilisée dans le TLFi.

- L'affichage à terme dans l'interface publique de la ou des catégories grammaticales de la vedette facilitera la compréhension et l'apprentissage du français.

Ce travail a été réalisé en deux étapes : de janvier à novembre 2022 puis de juin à novembre 2023.

Nous allons donc commencer par exposer les objectifs de ce projet et de la nécessité d'expliquer la démarche suivie. Ensuite nous parlerons des sources et enfin de la méthodologie adoptée pour finir, en conclusion, sur les possibilités de réutilisation.

Objectifs

La base de données était quasiment restée identique depuis le début des années 2000, à l'exception de l'ajout d'une date d'insertion/modification dans les tables en 2018-2019.

Les informations enregistrées sont assez simples comme nous pouvons le voir dans les figures ci-dessous : la table des graphies avec les entrées et leur identifiants, la table des synonymes avec les identifiants des entrées synonymiques et la table des antonymes avec les identifiants des entrées antonymiques.

La table des graphies :

id_graphie	int(5)	Identifiant
graphie	varchar(38)	Libellé
cnrtl	varchar(24)	Lien vers le dictionnaire CNRTL
date	timestamp	Date de création
nature	char(100)	Catégorie grammaticale

Tableau 1 : Table des graphies

La table des synonymes :

id	int(5)	Identifiant de la liaison
id_graphie1	int(5)	Identifiant renvoyant à la table des graphies
id_graphie2	int(5)	Identifiant renvoyant à la table des graphies
litigieux	tinyint(1)	1 : litigieux ; 0 : non litigieux (par défaut)
commentaire	int(10)	Identifiant renvoyant à la table des commentaires
date	timestamp	Date de création
observation	char(255)	commentaire interne
contexte	char(255)	phrase d'exemple

Tableau 2 : table des synonymes

La table des antonymes :

id_graphie1	int(5)	Identifiant renvoyant à la table des graphies
id_graphie2	int(5)	Idem avec comme contrainte d'être supérieur au précédent
date	timestamp	Date de création

Tableau 3 : table des antonymes

Enregistrer la catégorie grammaticale dans la base de données a souvent été un sujet récurrent au CRISCO sans avoir, jusqu'à présent, trouvé les sources adéquates.

Nous avons finalement pu récupérer celles utilisées dans le cadre du projet portant sur la polysémie évolutive provenant de l'ATILF.

Nous avons décidé en accord avec l'ATILF d'utiliser leurs fichiers afin de mémoriser dans la table des graphies (en ajoutant un nouveau champ intitulé « nature ») les différentes catégories grammaticales.

Il nous a paru important de rendre transparente la démarche d'enrichissement du DÉS avec la catégorie grammaticale pour fiabiliser le plus possible le résultat obtenu et justifier certains de nos choix parfois délicats. Le DÉS est depuis sa création un corpus régulièrement amélioré tous les mois avec les liens synonymiques, antonymiques et les variantes. Nous avons souhaité qu'il le reste avec cette donnée supplémentaire.

Méthode de constitution et/ou sources

Quatre sources différentes sous forme de fichiers en provenance de l'ATILF ont été utilisées (source 1a et source 1b, source 2a et source 2b).

Une source (source 1b) est issue du modèle IA, intitulé fr_dep_news_trf, utilisable avec la librairie Spacy en langage Python.

Les sources 1a et 1b ont été utilisées dans la première étape de traitement (de janvier à novembre 2022), les sources 2a et 2b dans la seconde étape de traitement (de juin à novembre 2023).

Source 1a

La source1a est un fichier tableur intitulé TLFI complet lemmes.xls de 54.280 lignes dont un extrait est présenté dans la table ci-dessous.

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.			

Tableau 4 : Extrait du tableur TLFI

Nous voyons que la seule colonne commune à toutes les lignes est la première avec le libellé de la graphie avec toutefois deux remarques importantes :

- la même graphie est parfois répétée et incrémentée d'un numéro (Ex : MEUBLE1, 2 ou 3).

- des parenthèses sont présentes, signalant des orthographes différentes autorisées

Ensuite, les lignes les plus simples sont celles avec uniquement une seconde colonne contenant la catégorie grammaticale.

D'autres lignes ont plusieurs colonnes avec des catégories grammaticales différentes, ou bien des extensions féminines (-AINE, -EUSE, etc.).

Source 1b

Le modèle fr_dep_news_trf est un pipeline entrainé et disponible en français contenant

un ensemble de composants : morphologiseur, analyseur syntaxique, régleur d'attributs, lemmatiseur,..

Ce modèle créé par la société Explosion (https://explosion.ai/) a été entraîné à partir de trois sources :

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

Le principe d'un entrainement supervisé utilisé dans ce cas pour créer un modèle consiste à sélectionner des données d'entrainement parmi ces 3 sources pour lesquelles les composants listés ci-dessus sont connus et à exécuter ce modèle sur des données d'évaluation pour vérifier les prédictions sur des exemples inédits, et à calculer le score de précision et ainsi de suite, jusqu'à l'obtention d'un modèle suffisamment fiable.

Ce modèle a été entrainé et est utilisable avec la librairie Spacy.

Nous avons fait ce choix car cette librairie est à la fois pertinente, unique et facilement utilisable avec le langage python.

Le composant qui nous intéresse est celui qui va associer une catégorie grammaticale aux mots. 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.

Source 2a

Dans le cadre du projet de modélisation graphique des notices historiques du TLFi, un programme a été créé pour extraire les données de 81 fichiers XML de l'ATILF et les enregistrer au format excel (xlsx). Nous sommes donc partis de ces 81 fichiers tableurs de ce projet pour en créer un unique de 49.854 lignes dont un extrait est donné ci-dessous.

Entrée	Catégorie grammaticale
absorber	verbe trans.
accenteur, accentueur	subst. masc.
despote	subst. masc. et adj.
déterminé, ée	part. passé, adj. et subst. masc.
dû, due	part. passé, adj. et subst. masc. sing.
narquois, -oise	adj. et subst. masc.
n'est-ce pas	loc. inv.
neuf1	adj. et subst. masc. inv.
neuf2, neuve	adj. et subst. masc.

Tableau 5 : Extrait des fichiers tableurs ATILF

Source 2b

Ce fichier plus récent de l'ATILF contient 103.328 lignes. Il est constitué de six colonnes : articleID, parentID, source, content, category, gender, feminine. La table ci-dessous donne quelques exemples d'entrées.

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	

Tableau 6 : extrait du fichier ATILF plus récent

On remarque que certaines entrées (colonne content) sont présentes sur plusieurs lignes, probablement liées à la notion d'acception. Cette notion d'acception est gérée différemment suivant les dictionnaires. Par exemple, pour accusé, le Grand Robert le présente sur une page en tant que nom et adjectif alors que le TLFi sur deux différentes.

Méthodologie

Les différentes étapes sont synthétisées dans le tableau ci-dessous et résumées dans l'ordre chronologique. La démarche globale utilisée est s'est inspirée de la méthode agile : en définissant concrètement pas à pas pour chacune des 2 étapes ce qui permettait de renseigner au mieux la base de données de façon la plus efficace. Nous avons donc commencé par utiliser les sources de l'ATILF à notre disposition pour renseigner le plus grand nombre d'entrées du DÉS possibles puis de compléter par des méthodes semi-automatiques pour les entrées restantes. Le détail est donné dans le document de travail (Chardon, 2020).

ÉTAPE	Type traitement	Détail du traitement	Résultat
N°1 : de janvier à novembre 2022	A partir de la source 1a de l'ATILF : TLFI complet lemmes.xls	Un total de 41.153 lignes avec 6981 entrées de type verbe , 4641 de type adjectif, 28588 de type substantif et 943 adverbes.	50 % de la base du DES renseignée (25.383 entrées sur 50.350 au total)
		5 traitements particuliers	65 % de la base renseignée (32.887 / 50.350)
	Sans source, traitements manuels avec le tableur avec consultation d'un dictionnaire en ligne	Recherche du schéma s' ou se un début du mot	68 % renseignée (34.298/50.350)
	A partir de la source 1b : Librairie fr_dep_news_trf avec spacy en langage Python	5 extractions en fonction de schémas : Mots finissant par -er, -ir, -é -dre, -eur, -tre, -ire, -oir, -ment ou commençant par s' ou à	83 % de la base renseignée (41.901 sur 50.496)
	Sans source, traitements manuels avec tableur consultation d'un dictionnaire en ligne	La catégorie grammaticale des 8488 entrées restantes a été corrigée manuellement par plusieurs personnes	100 % de la base renseignée
	Vérifications	Rechercher des incohérences sur les types grammaticaux Incompatibles entre 2 synonymes. Par exemple, un mot de type verbe synonyme d'un mot non verbe	250 entrées corrigées
N°2 : de juin à novembre 2023	A partir de la source 2a : 81 fichiers de l'ATILF au format xml récupérés sous forme d'un tableur de 49.854 lignes	Un programme de comparaison a permis d'extraire 662 entrées à vérifier.	211 entrées dans le DES corrigées
	A partir de la source 2b de l'ATILF avec 103.328 lignes	Un programme de comparaison permet de préciser le nombre d'entrées en commun avec le DES : 37.427. Parmi ces lignes, 336 sont à vérifier manuellement	237 entrées dans le DES corrigées
	Vérifications	Un programme amélioré par rapport à la phase 1 permet d'extraire 725 entrées à vérifier manuellement	318 entrées corrigées et 25 liaisons synonymiques supprimées

---

Tableau 7 : Etapes du traitement des fichiers ATILF

Première étape de janvier à novembre 2022

Traitement de la source 1a

Comme détaillé dans le document de travail, plusieurs étapes de traitements s'imposaient.

Tout d'abord, nous avons traité les entrées selon les catégories grammaticales : les verbes (6981), les adjectifs (4641), les substantifs (28.588), les adverbes (943).

Il faut signaler à ce moment une décision importante : une entrée présente sur plusieurs lignes finissant par des chiffres (comme MEUBLE) est considérée comme une acception c'est-à-dire ayant plusieurs sens ou plusieurs origines étymologiques. Les codes grammaticaux associés ont donc été séparés par un point virgule, de façon à les différencier des codes grammaticaux sur une seule ligne.

Par exemple, le champ nature de l'entrée MEUBLE dans le DÉS est substantif masculin ;

adjectif et substantif ; adjectif.

Cela correspond à :

- bien immobilier (subst. masc.)
- Qui se laboure ou se travaille facilement (adj.)
- Qui peut être transporté d'un lieu à un autre sans subir de détérioration (adj. et subst.)

Ces étapes ont permis de renseigner 50 % (25.383 sur 50.350) des entrées de la base du DÉS.

Ensuite, nous avons procédé à 3 traitements particuliers selon l'extension féminine en seconde colonne :

1. *-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.*

2. *-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.*

3. *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*

Les 2 traitements suivants ont demandé un travail manuel plus important. Nous avons traité 2619 entrées avec des orthographes différentes, les mots invariants, les prépositions, les interjections, les onomatopées en écartant les entrées de type « élément formant » (nyct-, oculi-, hodo-,..).

Puis enfin 1551 entrées de type locution à reformater pour être insérées automatiquement (par exemple : CATIMINI (EN) → en catimini, CONTREBORD (À) → à contrebord).

La procédure pour ces cinq traitements est détaillée dans les paragraphes « Introduire une première catégorie de mots mélangés » jusqu'à « Introduire une cinquième

catégorie de mots mélangés » du document de travail 1.

Traitement manuel sans source

Le traitement suivant est issu d'une constatation simple : parmi les 17.463 entrées dont la catégorie grammaticale n'est pas renseignée, 1.411 d'entre elles commencent par ou s'avèrent être des verbes.

Traitement de la source 1b

A partir des 16.052 entrées dans le DÉS qui n'ont pas de catégories grammaticales, nous avons gardé celles sans aucun espace soit 10.139 pour lesquelles la librairie Spacy apportait une réponse sur le code grammatical.

Ce résultat a été traité selon les catégories :

- Tout d'abord 588 entrées se terminant par -er et -ir avec le code « POS VERB » ont été vérifiées. Quelques corrections ont été réalisées comme décrottoir, débirentier ou parmentier.

- Ensuite, nous corrigeons les entrées finissant par « é » considérées à tort comme verbe et que nous avons notées comme participe passé. Puis les entrées avec le code PROPN (noms propres), PUNCT (ponctuations) ont été corrigées manuellement. Enfin celles avec le code NOUN et ADJ ont été sommairement vérifiées. Cela concerne un total de 951 lignes.

- Puis nous avons pris en compte les entrées avec des tirets et des apostrophes, nous récupérons ainsi 328 verbes commençant par « s' », 404 adverbes et 275 substantifs finissant par -ment.

- Pour les entrées commençant par « *à* », nous avons considéré que toute expression commencant ainsi est définie 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 la page wiktionary des locutions adjectivales en français ont été corrigées.

- La vérification de 4608 entrées retournées par Spacy comme étant « NOUN » :

    - 200 d'entre elles se terminant par er, ir et dre ont été vérifiées : 32 étaient

des verbes

- la vérification de 534 entrées se terminant par -eur, -ire et -oir n'a décelé qu'une erreur (stupéfaire : verbe)

- les entrées restantes sont restées des substantifs

Traitement manuel sans source

Les 8488 entrée du DÉS sans catégorie grammaticale ont été vérifiées par plusieurs personnes selon plusieurs types de filtrage détaillés dans le document de travail sus-cité.

Vérifications

Il nous a semblé intéressant de mettre en place des tests pouvant potentiellement faire apparaître des incohérences. Par exemple, si une entrée de type verbe sans être substantif, ni adjectif, ni adverbe, ni locution, est synonyme d'une entrée qui n'est ni un verbe, ni un adverbe ni une locution, alors une vérification s'imposait. L'ensemble des tests est donné dans la table 9 du document de travail.

250 entrées ont été corrigées.

Seconde étape de juin à novembre 2023

Cette seconde phase a permis, non pas de renseigner les catégories grammaticales des entrées du DÉS puisqu'elles l'étaient toutes, mais plutôt de comparer le DÉS avec ces deux sources afin de corriger et de compléter les catégories grammaticales.

Traitement de la source 2a

Nous avons cherché tout d'abord à calculer des indicateurs généraux pour comparer les 2 sources :

- 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 catégories grammaticales du DÉS incluses dans TLFI (code 2) : 449
- Nombre d'entrées en commun avec les catégories grammaticales 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

Suite au traitement des entrées différentes selon les codes 2 et 3, nous avons conclu que les catégories grammaticales de :

- 91 entrées (code 2) et 120 (code 3) étaient à corriger manuellement
- 358 entrées (code 2) étaient dues à une codification différentes pour les verbes
- 93 entrées (code 3) étaient dues à des inversions (par exemple, "adj. et subst." d’un coté et "subst. et adj." de l’autre)

Traitement de la source 2b

Nous avons comme pour la source précédente calculé des indicateurs :

- Nombre d’entrées dans le TLFi2 : 103.328
- Nombre d’entrées uniques 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

Sur le 37.427 entrées communes, nous avons réalisé un traitement pour ne retenir que les entrées à vérifier.

L’algorithme du traitement est le suivant :

```
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
```

Figure 1 : Algorithme du traitement

Les 336 entrées pour lesquelles la colonne « ok ? » est à False ont été vérifiées

manuellement et corrigées dans la base du DÉS.

Vérifications

Pour ces dernières vérifications, nous sommes 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 n'effectuons pas de vérification.

Cela donnait toutefois 5.828 lignes à revoir, ce qui était impossible à vérifier à la main. Nous avons donc exclu des liaisons qui vérifiaient les critères de la table 15 du document de travail (par exemple nous avons écarté le cas où le premier mot est un adjectif et le second un substantif).

Les 725 lignes restantes ont été vérifiées manuellement (voir le fichier CatGramErreursAcceptionsRecup_2023-07-13.csv sur le git).

Présentation du contenu et de l'organisation du jeu de données

Le champ nature des 50.000 entrées du DÉS est renseigné avec une des 476 combinaisons de codes grammaticaux enregistrés. L'ensemble de ses combinaisons est présent sur le git public (https://git.unicaen.fr/crisco-des-public/descatgram/-/blob/master/catgram_20240411.csv) : nous en avons répertorié 476. La table ci-dessous en donne un extrait :

---

catgram_20240411

Combinaisons de codes grammaticaux enregistrées	Nombre dentrées concernées
adj.	4426
adj.;adj. et pron. indéf.	1
adj.;adj. et subst.	15
adj.;adj. et subst. masc.	2
adj.;adj. masc.	1
adj.;adv.	4
adj.,adv. et prép.	1
adj.,adv. et subst.	9
adj.,adv. et subst. masc.	2
adj.,adv. et subst. ;subst. masc.	1
adj.,adv.,prép. et subst. masc.	1
adj.;adv.;subst. masc.	4
adj. composé	1
adj. et adv.	9
adj. et adv.;adj. inv.	1
adj. et adv.;subst. fém.	1
adj. et adv.;subst. masc.	2
adj. et interj.	2
adj. et pron.	1
adj. et pron. indéf.	3
adj. et pron. indéf. plur.	1
adj. et pron. indéf. subst.	1
adj. et subst.	1565
adj. et subst.;adj.;adv.;subst. masc.	1
adj. et subst.;adv. et subst.;adj.	1
adj. et subst.,adv.;subst. masc.	1
adj. et subst. fém.	114

Tableau 8 : Extrait du champ nature du DÉS

Modalités d’accès aux données

L’ensemble des données du DÉS sont téléchargeables sur la plate-forme ORTOLANG pour la communauté scientifique (authentification requise) à partir de ce lien direct : (https://hdl.handle.net/11403/des/v3). Il existe un git public (cf. supra).

L’exploitation des données : analyse et interprétations

L’ajout de cette information supplémentaire va permettre de filtrer selon :

- les catégories grammaticales
- les entrées ayant plusieurs sens par la présence d’un point virgule séparant les codes (acceptions)
- et de faciliter des recherches spéciales comme les entrées étant à la fois verbes et substantifs ou toutes les locutions, etc...

Perspectives de réutilisation

L'ajout de la catégorie grammaticale apporte plusieurs bénéfices :

Tout d'abord, dans le domaine de recherche développée au CRISCO sur les graphes, il sera possible d'affiner des extractions afin d'initier ou de poursuivre des recherches :

- L'étude de Bernard Victorri « Quand les mots s'organisent en réseaux » qui étudie l'ensemble des verbes français (9000 au total) qui s'organisent sous forme de graphes selon 4 axes : verbes de fuite et rejet, de production et croissance, de lien et communication, de destruction et de dégradation (Victorri, 2010)

- L'ouvrage de Fabienne Venant « Représentation et calcul dynamique du sens » (Venant, 2010) dans lequel plusieurs chapitres sont consacrés à l'étude globale des adjectifs (3699 exactement) sous forme de graphes de synonymie

De plus, les données étant déposés avec la licence Creative Commons sur la plate-forme ORTOLANG, elles peuvent être utilisées dans d'autres domaines de recherche ; ainsi le dépôt réalisé en novembre 2022 a été consulté plusieurs centaines de fois et téléchargé plus d'une centaine de fois.

Il est aussi possible d'envisager la comparaison de cette base avec d'autres lexiques.

Un premier exemple concerne le Réseau Lexical Français contenant 28.059 entrées, avec leur catégorie grammaticale et les liens paradigmatiques (relation de synonymie) et les liens syntagmatiques (qui reflètent les mots très souvent employés les uns avec les autre)s. Il serait intéressant de comparer les 2 corpus et de les enrichir ou de les corriger mutuellement.

Un second exemple, Holinet est le corpus associé au projet JeuxDeMots (JDM) qui vise à construire un réseau lexical et sémantique du Français. Il contient plusieurs types de relation (61 d'après cette page) dont les relations de synonymie et d'antonymie et des informations de grammaire.

Enfin, la possibilité de l'afficher dans l'interface graphique d'interrogation du DÉS facilitera l'apprentissage des apprenants. En effet, les différents sens d'un mot sont liées dans beaucoup de cas aux catégories grammaticales. Nous avons donné l'exemple de « mousse » dans l'introduction. Il y a également « trouble » qui est adjectif (qui n'est pas limpide, ambigü), substantif masculin (une agitation) et substantif féminin (petit filet de pêche) ou « critique » qui est un adjectif (qui comporte un danger), substantif féminin (un jugement), substantif masculin (une personne qui a le pouvoir de juger).

Actuellement, les synonymes d'une entrée sont présentés dans l'ordre alphabétique sans organisation par sens. Réorganiser l'interface web du DÉS en ajoutant la catégorie grammaticale des entrées pourra faciliter l'apprentissage des apprenants en leur

présentant, par exemple, les synonymes par catégorie grammaticale.

Références bibliographiques

Chardon, L. (2024). *Insertion des catégories grammaticales dans le Dictionnaire Électronique des Synonymes (DES)*. Document de travail. URL : halshs-03956407v2

Chardon, L. (2020). L'espace sémantique du "Dictionnaire électronique des synonymes" (DES) et les méthodes de regroupement de sens : l'exemple de "sec". *Syntaxe et Sémantique*, *1*(21), 87-126. URL : 10.3917/ss.021.0087 ; halshs-03155459

Ploux, S., Victorri, B. (1998). Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes. *Revue TAL : traitement automatique des langues*, 39, 61-182. URL : halshs-00009433

Ploux, S. (1995). *Traitement des synonymes*. CNRS, Université de Caen. URL : hal-02430301

Ploux, S. (1996). *Une étude pour le traitement informatique de la synonymie*. Document de travail. URL : hal-02430342

Victorri, B., Manguin, J.L. (1999). *Représentation géométrique d'un paradigme lexical* [Communication orale]. Conférence TALN 1999, Cargese (Corse), France. URL : hal-04520029

Victorri, B. (2010). Quand les mots s'organisent en réseaux. *L'Archicube*, 8, 53-59. URL : halshs-00666584.

Venant, F. (2010). *Représentation et calcul dynamique du sens.* Editions universitaires européennes. URL : hal-04526033.

Notes

1 Le dépôt git est consultable à l'adresse suivante : https://git.unicaen.fr/crisco-des-public/descatgram/.

2 L'espace sémantique d'un mot appelé vedette est une représentation graphique qui va illustrer la portée de chacun de ses sens. Il est calculé à partir du calcul des cliques et d'une réduction de dimension. Par exemple ce tutoriel sur l'espace sémantique de curieux va donner ses différents sens : (bizarre, étonnant,...) (fureteur, indiscret ...), spectateur, (indiscret, désireux..)

3 Les atlas sémantiques de Sabine Ploux sont un modèle de représentation géométrique du sens des mots testé pour sa pertinence cognitive. Voir https://www.atlas-semantiques.eu/Apropos_as.html?l=FR/.

4 La collection HAL du DÉS est accessible par ce lien : https://hal.science/CRISCO-DES/.