NEOPHILOLOGICA 2023, T. 35, 1–19
ISSN 2353-088X
DOI: https://doi.org/10.31261/NEO.2023.35.10

Morfetik – mises à jour et évolutions d'une ressource en ligne

Aude Grezka
CNRS, LIPN
Université Sorbonne Paris Nord
France
https://orcid.org/0000-0002-4582-3428

Morfetik – updates and upgrades to an online resource

Abstract
In this article, a morphological linguistic resource for contemporary French called Morfetik is presented. The evolution of the resource and its various linguistic and technological characteristics are discussed. Additionally, an overview of the numerous tools integrated into the resource is provided. Morfetik represents a continuously evolving platform designed to progress and enhance the processing of textual data.

Keywords
Morphological linguistic resource, Morfetik, NLP, language variation, French

1. Introduction

La création d'une ressource lexicale ne se résume pas à un listage d'éléments lexicaux mais dans une définition rigoureuse et actualisable du lexique et de ses diverses flexions. Ces ressources morphologiques sont utilisées dans divers domaines de la linguistique et de la recherche en Traitement automatique du langage (TAL). Elles sont essentielles pour l'analyse et la génération de mots, la lemmatisation, l'étude des relations entre les mots, la modélisation linguistique et bien d'autres applications linguistiques.

Dans cette perspective, est né le projet Morfetik. Ce projet a débuté en 2008[1] avec l'informatisation de la ressource lexicale de M. Mathieu-Colas (M. Mathieu-Colas, 2009 ; P.-A. Buvet *et al.*, 2009) et de sa structuration, travail de plus d'une vingtaine d'années de collecte et de description. En 2015, A. Grezka a repris la direction du projet et a repensé la ressource avec les technologies actuelles[2]. La base de données a également été actualisée et enrichie considérablement, notamment avec l'ajout des unités polylexicales, des verbes pronominaux et des néologismes, totalement absents auparavant (A. Grezka, 2017 ; 2020). Morfetik est une ressource en ligne en accès libre[3] qui présente les caractéristiques suivantes : une large couverture, des informations précises et fiables, le respect des normes et une évolutivité garantie. La ressource lexicale Morfetik est un dictionnaire morphologique des unités lexicales simples et polylexicales du français contemporains (noms, adjectifs, déterminants, pronoms, verbes, adverbes, prépositions, conjonctions, interjections, locutions, etc.). Il est ainsi possible d'obtenir toutes les formes associées à n'importe quel mot français, que ce soit le pluriel des noms, le féminin et le pluriel des adjectifs, ou encore les formes conjuguées des verbes, et bien d'autres. De même, Morfetik permet d'identifier le mot de base, appelé « lemme », correspondant à n'importe quelle forme fléchie.

Les données sont organisées en tables et constituent le point de départ d'un système de traitement qui comprend un moteur de flexion, un dictionnaire des formes fléchies, des interfaces de consultation et d'interrogation, ainsi qu'une interface d'édition réservée aux éditeurs pour mettre à jour la base de données (assurer la maintenance et l'exploitation de la ressource). Grâce à ce système, il est possible de générer automatiquement toutes les formes simples et complexes de la langue française et d'apporter également des informations sémantiques lorsque cela est nécessaire (domaines, par exemple), de contexte (analyse et suivi du mot dans la presse), de fréquence, et bien d'autres éléments. Depuis quelques années Morfetik s'est imposé comme un outil linguistique essentiel, offrant des analyses morphologiques précises et des informations linguistiques approfondies.

L'article présente la ressource par le biais de ses évaluations linguistiques et informatiques. Nous proposons tout d'abord de revenir brièvement sur le fonctionnement général de Morfetik. Puis, nous présentons les différentes nouveautés. Nous verrons notamment les outils qui ont été ajoutés à la ressource. Mor-

[1] Le projet a débuté au laboratoire LDI (Lexiques, Dictionnaires, Informatique, UMR 7187 CNRS), anciennement LLI (Laboratoire de Linguistique Informatique) dirigé par Gaston Gross.

[2] Ce travail bénéficie partiellement d'une aide de l'IdEx Université Paris Cité (ANR-18-IDEX-0001) au titre du Labex Empirical Foundations of Linguistics – EFL.

[3] Lien vers la ressource (version 0.2(2022)) : https://tal.lipn.univ-paris13.fr/morfetik.

fetik a pour vocation de s'enrichir progressivement pour améliorer la chaîne de traitement des données textuelles.

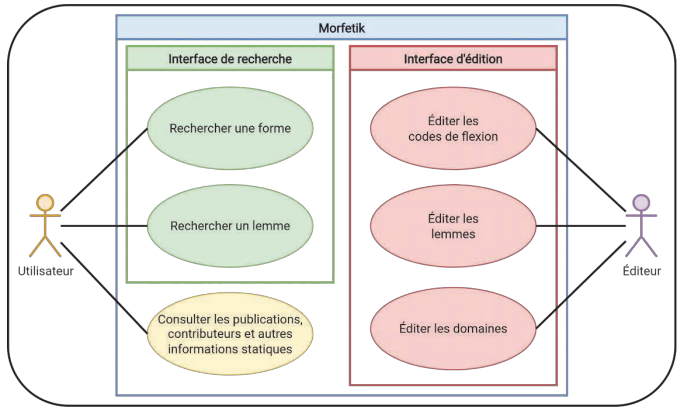
2. Fonctionnement de Morfetik

La plateforme Morfetik est découpée en trois grands blocs : l'interface de recherche de la base de données, la fonction principale de Morfetik ; des pages d'informations (page d'accueil, publications, etc.) et une interface d'édition accessible uniquement par les éditeurs pour permettre la mise à jour de la base de données. Dans cette partie, nous détaillons l'architecture générale du système et ses principaux modules. Du point de vue technique, l'un des plus gros travail que nous avons dû réaliser, a été la reconstruction du moteur de flexion et la création d'une interface éditeur pour faciliter le travail collaboratif.

2.1. Cas d'utilisation

Morfetik à deux cas principaux d'utilisation : la recherche par mot (lemme) et la recherche par forme. À ces cas s'ajoute la gestion des données, avec une interface d'édition pour que les éditeurs de la ressource puissent ajouter des mots ou des informations plus facilement :

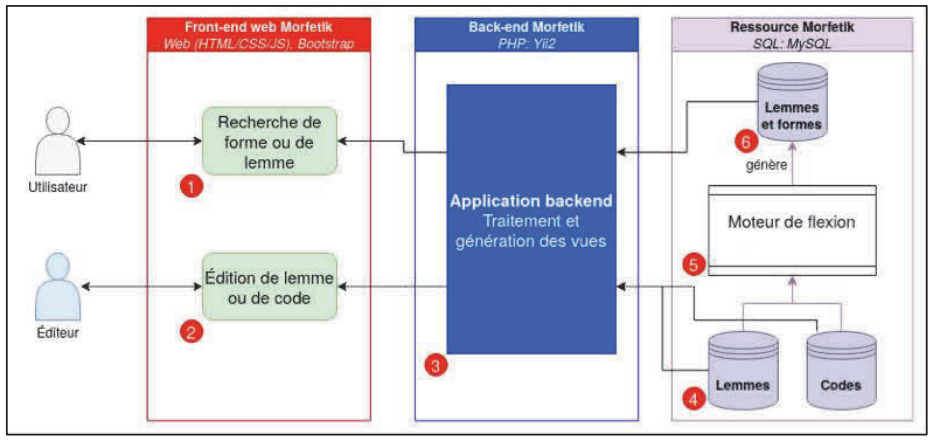
Figure 1
Diagramme des cas d'utilisation de Morfetik (v.0.2)



2.2. Architecture de Morfetik

L'architecture de Morfetik est la suivante :

Figure 2
Architecture de Morfetik (v.0.2)



1. Les utilisateurs peuvent rechercher une forme ou un lemme.
2. Les éditeurs peuvent modifier les lemmes ou les codes.
3. Dans les deux cas (utilisateur et éditeur), le back-end[4] va s'occuper de récupérer les informations.
4. Dans le cas d'une demande d'édition, le moteur va permettre de modifier les lemmes et les codes de la ressource Morfetik.
5. Du côté de la ressource, le moteur de flexion, écrit en SQL, s'occupe de générer la table des formes et des lemmes. Le moteur effectue la flexion des lemmes : il permet de passer d'un lemme et d'un code à des formes.
6. Cette table est utilisée pour les requêtes de formes et de lemme. L'application back-end effectue de la logique simple pour récupérer les résultats.

Le choix de cette architecture, d'inclure le moteur de flexion non pas dans l'application back-end mais dans la base de données elle-même, présente un avantage. La valeur de Morfetik réside dans les données linguistiques et non dans ses interfaces d'utilisation. Le moteur de flexion est primordial au fonctionnement de Morfetik. Ainsi, si le front-end[5] ou encore le framework[6] utilisés viennent

[4] La partie invisible de l'application web (côté serveur).
[5] La partie visible de l'application web (côté client).
[6] Ensemble d'outils et de composants logiciels à la base d'une application.

à changer, l'accès aux données, le plus important, restera toujours présent, rendant ainsi Morfetik modulaire. L'objectif dans la conception de Morfetik est de créer une application qui puisse durer dans le temps. Les frameworks, le back-end ou le front-end viendront peut-être à changer mais la base ne changera pas.

2.3. Reconstruction et mise en fonctionnement du moteur de flexion

L'un des problèmes majeurs et techniques que nous avons dû régler à la reprise du projet fut la reconstruction du moteur de flexion. Depuis de nombreuses années, Morfetik était statique puisque le générateur automatique de formes fléchies ne fonctionnait plus. Le système permet maintenant de générer automatiquement l'ensemble des formes simples et composées du français et donc de mettre à jour les données.

Les formes sont générées automatiquement à partir d'une table des lemmes et d'une table de modèles morphologiques représentés par des codes de flexion (pour les mots invariables, comme les adverbes, un simple listage suffit). Les codes (environ 400 actuellement) représentent les formes que vont prendre chaque lemme dans des situations différentes : temps, masculin ou singulier, pluriel des noms[7], féminin/pluriel des adjectifs, etc. Le générateur de formes fléchies est la composante principale du système qui prend en charge l'interprétation des tables pour générer le lexique des formes fléchies (c'est-à-dire le résultat de la recherche). À chaque forme sont associés un lemme ainsi que différentes informations de catégorie, de genre et de nombre (noms, adjectifs) ou de temps, de mode, de personne et de genre (verbes). Il comprend un moteur principal ainsi que des modules. Le moteur principal se charge de la collecte des informations

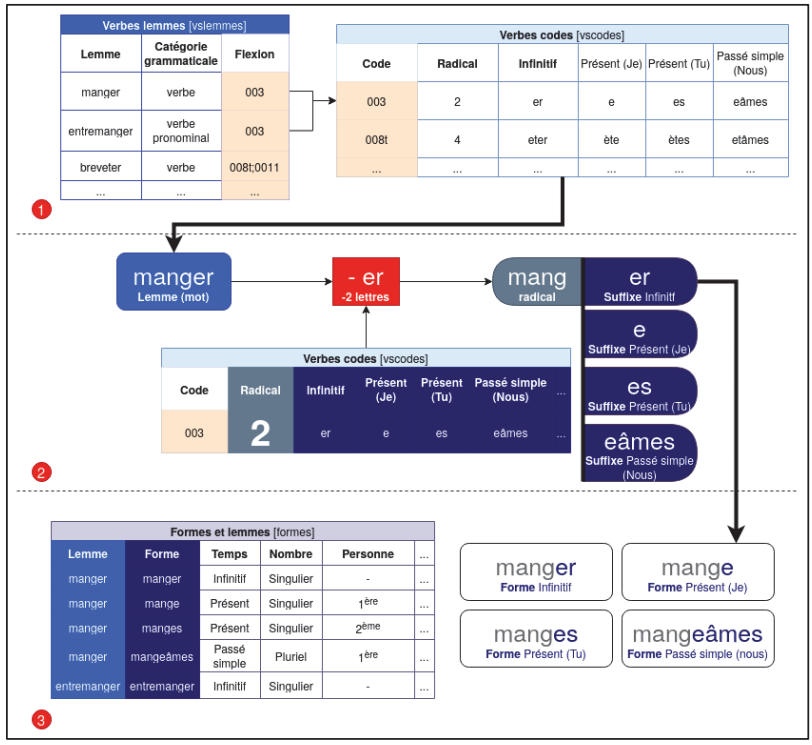
[7] Pour des raisons à la fois théoriques et pratiques, les noms n'ont pas été fléchis en genre, à la différence d'autres dictionnaires, il n'y a donc dans la nomenclature que des noms masculins et des noms féminins (M. Mathieu-Colas, 2009). Du point de vue théorique, le genre n'a pas le même statut pour les noms et les adjectifs. Il est inhérent au nom, d'où la variété des formes : à côté de noms épicènes (*un élève, une élève*), toutes sortes de substitutions ou de suffixes peuvent apparaître (*un loup, une louve*). Il y a même une rupture morphologique entre les deux genres (*un dieu, une déesse ; un roi, une reine ; un garçon, une fille*). Du point de vue pratique, le traitement du féminin en termes de flexion pose de nombreuses difficultés, dues à la polysémie. Si l'on veut regrouper en une seule entrée *maître* et *maîtresse*, il faut préciser que cette correspondance, usuelle dans le domaine scolaire (« le maître / la maîtresse a dit… »), ne vaut pas pour tous les emplois : ainsi, il est des cas où *maître* ne peut être qu'au masculin (« les grands maîtres de la peinture » vs « *les grandes maîtresses de la peinture »), ou à l'inverse au féminin (« il a rejoint sa maîtresse » vs « *elle a rejoint son maître »). Même des formes apparemment simples peuvent être problématiques : par exemple, si *boulanger* désigne toujours quelqu'un qui fabrique et qui, le cas échéant, vend du pain, *boulangère* peut aussi signifier « femme d'un boulanger » (Larousse, TLFI).

dans les tables, de la création des formes fléchies à partir du lemme et du code de flexion approprié, de l'attribution des catégories grammaticales correspondantes (genre, nombre, etc.) et des autres métadonnées, et enfin il centralise l'exécution des modules. Ces modules sont de plusieurs ordres. Chaque partie du discours nécessitant une flexion dispose d'un module spécifique permettant de gérer finement les particularités des tables qui lui correspondent, comme les différents types de codes pour la rareté, ainsi que les différents moyens de signaler les formes défectives ou les variantes. Cette architecture modulaire permet de pouvoir assez facilement envisager la réutilisation du moteur pour d'autres langues ou d'autres ressources.

L'intérêt principal de ce système, est qu'au lieu de rentrer toutes les flexions, toutes les formes d'un mot, d'un verbe ou d'une expression, les éditeurs n'ont besoin de rentrer que le lemme dans la base de données en lui associant un ou des codes de flexion qui vont ainsi permettre de créer automatiquement la forme fléchie. Le schéma suivant permet de comprendre le moteur de flexion :

Figure 3

Fonctionnement du moteur de flexion pour un verbe



Le processus de flexion se fait en trois étapes :

1. Chaque lemme possède une catégorie grammaticale et une flexion, correspondant à un code. Un lemme peut posséder plusieurs flexions. Le moteur de flexion identifie tous les codes correspondants au lemme.
2. Chaque code possède un champ « Radical », indiquant le nombre de lettres à enlever pour avoir le radical du mot. Le moteur de flexion associe ensuite le radical à chaque suffixe.
3. Des formes fléchies sont créées. Le moteur de flexion joint les résultats dans une table, avec des informations syntaxiques, comme le temps et le nombre si le lemme est un verbe, ou le genre si le lemme est un nom.

Prenons l'exemple d'un mot simple qui n'a que 2 formes fléchies : *sandwich*. Au pluriel, ce nom masculin peut prendre la forme *sandwichs* ou la forme *sandwiches*. Pour que le moteur de flexion puisse créer toutes les formes de ce mot, il a été codé avec les codes 01 (qui ajoute « s » au pluriel) et avec le code 25 (qui ajoute « es » au pluriel). Le moteur de flexion va donc créer les 2 formes avec le code 01 (un *sandwich,* des *sandwichs*) puis les 2 formes avec le code 25 (un *sandwich,* des *sandwiches*). Pour éviter d'avoir des doublons dans la base de données, lorsqu'une forme existe déjà dans la base, on ne la crée pas. Ainsi dans cet exemple, le moteur de flexion ne crée que 3 formes (un *sandwich,* des *sandwichs/sandwichs*) puisque le singulier est identique malgré 2 codes différents. La majorité des codes ont également pour fonction de retirer une partie du lemme pour obtenir le radical avant d'ajouter la terminaison (par exemple le code 03 des noms retire la dernière lettre et ajoute « -ux » au pluriel, on peut le voir sur le mot *idéal* → *idéaux*).

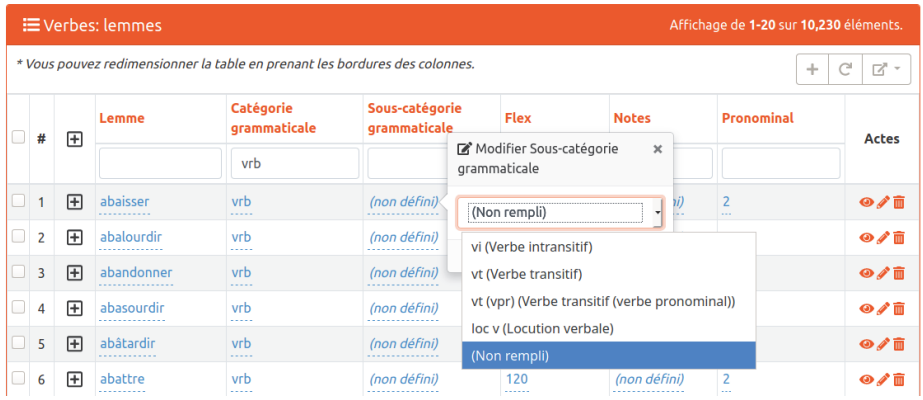
Chaque code est associé à plusieurs lemmes, et un lemme est associé à un seul code.

2.4. Création d'une interface éditeur

Dans les premières versions de Morfetik, les éditeurs modifiaient la base de données directement. Un de nos objectifs dans ce projet a été que la ressource soit facile d'utilisation pour les utilisateurs comme pour les éditeurs. Nous avons donc créé une interface CRUD (Create Read Update Delete : interface permettant d'interagir avec des données, on peut consulter, créer, mettre à jour et supprimer des données) pour interagir directement avec Morfetik :

Figure 4

Édition directe de la table avec menu déroulant (image prise sur le site Morfetik)



L'interface d'édition permet maintenant d'éditer la base Morfetik et sa structuration, de rechercher, modifier ou bien ajouter de nouveaux mots. L'interface d'édition possède également des menus déroulants qui sont modifiables directement. La grille d'édition permet l'édition de fiche, la suppression d'un lemme ou la suppression de masse. La recherche est directement intégrée à la grille. On peut également exporter en format TXT ou CSV, pour permettre le traitement par Excel.

L'interface éditeur permet ainsi aux utilisateurs de modifier ou de personnaliser la ressource de manière conviviale et intuitive, sans avoir à manipuler directement le code ou les paramètres techniques. Cela ouvre la porte à un plus large éventail d'utilisateurs, même ceux qui ne possèdent pas de connaissances approfondies en programmation. Enfin, l'utilisation de l'interface éditeur permet de réduire les erreurs potentielles. Les utilisateurs peuvent effectuer des modifications dans un environnement contrôlé et guidé, ce qui diminue les risques d'introduire des erreurs de syntaxe ou d'autres problèmes liés à la manipulation directe du code.

3. Caractéristiques linguistiques de la ressource

Nous présentons ici les caractéristiques linguistiques de la ressource, ainsi que les nouveautés (veille linguistique, implémentation des unités polylexicales, des domaines et des verbes pronominaux).

3.1. Large couverture et fiabilité des données lexicographiques

Morfetik est actuellement la ressource morphologique du français la plus exhaustive. La ressource a en effet été comparée avec des ressources lexicales analogues en français[8] (A. Grezka *et al.*, 2015). La couverture lexicale a également été validée par comparaison avec trois corpus du français suffisamment volumineux et représentatifs de la langue générale (les 10 ans du Monde, le Wikipédia français et la version française de Wacky). La ressource contient plus de 1.000.000 de formes.

L'une des spécificités de Morfetik réside dans la nature des sources. Plutôt que de faire table rase de l'héritage lexicographique en s'en remettant exclusivement à des procédures d'extraction automatique, il nous a semblé préférable de nous appuyer, dans un premier temps, sur la richesse des données consignées dans les dictionnaires, qui ont le double mérite d'exister et d'être particulièrement fiables. Toutes les variantes graphiques attestées dans les ouvrages ont été retenues. En effet, l'inclusion des variantes graphiques (A. Grezka, 2020) dans la base de données vise à fournir une ressource complète, précise et respectueuse de la diversité linguistique, tout en répondant aux besoins des différents utilisateurs de la langue, qu'ils soient novices ou experts, en leur donnant une vue complète des possibilités orthographiques. Les principales sources consultées ont été les suivantes :

- le *Petit* et le *Grand Robert,* le *Petit Larousse illustré* et le *Lexis* ;
- le *Trésor de la langue française* (avec toutefois un filtrage des entrées, pour exclure certaines données trop datées ;
- le *Harrap's* et le *Robert & Collins* ;
- des dictionnaires d'argot (le *Dictionnaire de l'argot de Larousse*) ;
- des tables de conjugaison (dont le *Bescherelle*) ;
- *Le Bon Usage* de GREVISSE et des dictionnaires de « difficultés » pour le traitement des cas problématiques.

Nous avons également pris en compte le DELAS (*Dictionnaire électronique du LADL*, cf. B. Courtois, 1990), avec, ici encore, un filtrage des données, afin d'éliminer certains artefacts.

Afin d'amorcer l'intégration des langues de spécialité au sein de la base, nous avons mis à contribution, dans une large mesure, le *Grand Dictionnaire encyclopédique Larousse* (GDEL), et exploité systématiquement, pour deux domaines particuliers – la médecine et la minéralogie –, des dictionnaires spécialisés.

---

[8] Notamment avec les ressources lexicales : GLAFF (Hathout *et al.*, 2014 ; Sajous *et al.*, 2013, 2014), Lefff (Clément *et al.*, 2004 ; Sagot, 2010), Morphalou et Dicolecte.

Ce noyau historique de Morfetik étant constitué, nous avons pu, dans un deuxième temps, entreprendre l'actualisation et l'élargissement des données, sur plusieurs plans simultanés :

– l'intégration des néologismes, des nouvelles graphies (cf. les Rectifications de l'orthographe 1990) et des nouveaux féminins ;
– l'exploration d'autres spécialités ;
– l'exploitation des grands corpus.

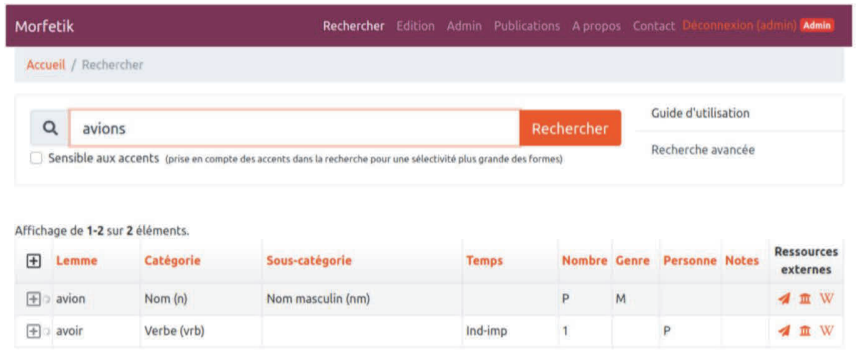
3.2. Mises à jour régulières : veille linguistique

Cet inventaire n'est pas définitif et continue à évoluer. Il était en effet important pour nous d'effectuer une veille linguistique, qui ne se faisait plus sur Morfetik. Les dictionnaires, comme toutes les ressources linguistiques, nécessitent une confrontation continuelle avec des corpus, pour surveiller, collecter des informations sur les évolutions, les tendances et les nouveautés dans le domaine de la langue. Cette veille est essentielle pour rester à jour dans un monde linguistique en constante évolution.

Ainsi, nous actualisons quotidiennement la base de données par l'ajout de néologismes et l'intégration de nouvelles spécialités, rendu possible par les programmes liés à Morfetik, grâce à la connexion que nous avons établie avec les ressources Néoveille (https://tal.lipn.univ-paris13.fr/neoveille) et France Terme (https://www.culture.fr/franceterme)[9] :

Figure 5

Interface de recherche de Morfetik (image prise sur le site Morfetik)



[9] Cette partie du projet a obtenu des subventions de la DGLFLF entre 2015 et 2019. La DGLFLF élabore la politique linguistique du Gouvernement en liaison avec les autres départements ministériels.

La plateforme Néoveille (E. Cartier, 2011, 2019 ; E. Cartier & J.-F. Sablayrolles, 2009) vise à détecter automatiquement, décrire linguistiquement et suivre l'évolution des innovations lexicales en corpus dynamique. Sa connexion avec Morfetik permet une mise en parallèle permanente avec des corpus afin de suivre l'évolution fréquentielle des lexies et de repérer ainsi celles qui semblent s'implanter (c'est-à-dire des néologismes qui atteignent une fréquence suffisante, sur une période donnée) pour les intégrer à Morfetik.

La ressource FranceTerme, quant à elle, est une base de données terminologiques de la Délégation Générale à la Langue Française et aux Langues de France (DGLFLF) du ministère de la Culture, qui rassemble les récents néologismes spécialisés et avalisés par la Commission d'enrichissement de la langue française et publiés au Journal officiel, remplaçant les termes importés d'autres langues. La base terminologique comporte plus de 8.500 termes français dans différents domaines scientifiques et techniques. Elle a pour mission de promouvoir l'utilisation de mots en français, pour enrichir la langue française et éviter ainsi son recul dans le monde. Ainsi, pour désigner les réalités nouvelles qui se créent constamment, des termes français sont recommandés par des spécialistes. Par exemple, le domaine de l'informatique est notoirement connu pour user d'anglicismes, alors même que des équivalents français existent parfois. Ces termes font donc leur entrée dans Morfetik : *plateforme de données* (pour *data hub*), *apprentissage profond* (pour *deep learning*), *banque de données* (pour *data bank*), etc.

Ces ressources nous permettent donc de compléter les lacunes et d'enrichir la terminologie qui ne cesse d'évoluer, en particulier dans les langues de spécialités (nouvelles informations, nouveaux concepts ou nouvelles découvertes).

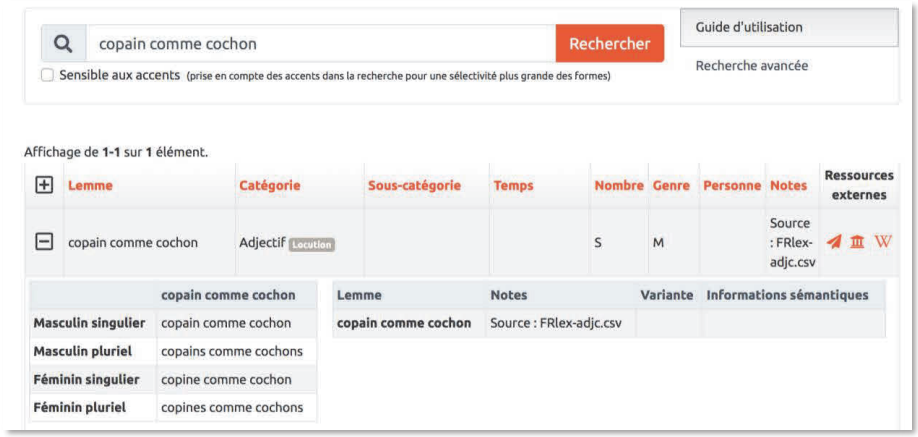
3.3. Implémentation des unités polylexicales

Aucun traitement informatisé n'est concevable si les unités complexes ne sont pas clairement identifiées, ce qui suppose un recensement très large et une description minutieuse. Nous avons, en 2022, implémenté dans Morfetik les unités polylexicales, quel que soit leur degré de figement (mots composés, mots à trait d'union, locutions, etc.)[10] :

[10] Nous ne faisons pas de différence, ici, entre les diverses appellations des séquences figées : noms ou adjectifs composés, locutions nominales ou adjectivales, noms ou adjectifs polylexicaux, etc. On conviendra de parler d'« unités polylexicales » dans un sens indifférencié pour désigner tous les syntagmes qui comportent, à quelque titre que ce soit, une part de figement.

Figure 6

Recherche d'une expression (image prise sur le site Morfetik)



Sur le même principe que pour les mots simples, un dictionnaire des unités complexes a été construit (M. Mathieu-Colas, 2014 ; G. Gross, 1996). La tâche n'était pas simple, compte tenu de la diversité des formes (plus de 700 types morphologiques pour les seuls noms composés). Il fallait compter aussi avec la complexité de certains schémas flexionnels, notamment pour les mots à trait d'union, qui comportent de nombreuses exceptions. Prenons l'exemple de l'adjectif composé *franc-comtois* : on trouvera au masculin pluriel *FRANCS-comtois* (des *fromages FRANCS-comtois*) mais au féminin pluriel *FRANC-comtoises* (des *horloges FRANC-comtoises*). L'autre difficulté rencontrée lors de l'inventaire, est que l'orthographe des mots composés est loin d'être bien définie dans les dictionnaire traditionnels : un nombre non négligeable d'entrées sont écrites soudées dans certains dictionnaires, avec un trait d'union dans d'autres ; ou parfois l'information de flexion n'apparaît pas.

> Les difficultés de l'orthographe française ne tiennent pas seulement à la complexité des règles traditionnelles et à la multitude des exceptions, elles résultent également, de manière plus insidieuse, de l'état d'« incertitude » qui caractérise une partie du lexique. La norme n'est pas si bien fixée qu'on se plaît à le croire : un certain nombre de variantes existent, qui désignent indirectement les failles de réécriture et invitent par là même à s'interroger sur la logique du système et les fondements de l'usage.
>
> (M. Mathieu-Colas, 1990 : 104)

Nous avons donc adopté un système suffisamment flexible pour pouvoir prendre en charge les différents cas de figure.

Pour les noms, le codage morphologique des unités polylexicales repose sur les mêmes principes fondamentaux que celui des noms simples, mais la complexité des formes implique quelques ajustements. La difficulté résulte principalement dans le fait que la flexion peut être terminales et/ou internes, multiples ou être absente :

- absence de flexion : le mot n'a qu'une forme, par exemple le nom *beaux-parents* (nms) n'existe qu'au pluriel ; le *bien-être*, uniquement au singulier ;
- flexion terminale : le mot peut se rencontrer au singulier ou au pluriel, mais la flexion fonctionne comme celle des mots simples : une *demi-heure*, des *demi-heureS* ;
- flexion interne : le nom peut avoir une ou plusieurs flexions internes : un *arrière-grand-père*, des *arrière-grandS-pèreS* ; une *assiette à soupe*, des *assietteS à soupe*. La difficulté réside dans la gestion des variantes flexionnelles puisqu'un même élément peut donner lieu à deux flexions distinctes : *match aller matchS* ou *matchES aller* ou alors un même élément peut être fléchi ou non fléchi : un *haricot vert sans fil*, des *haricotS vertS sans fil(S)* ;
- flexion interne et terminale : une même unité peut relever simultanément de deux types de flexion : un *social-démocrate*, des *social-démocrateS*, des soci*AUX-démocrateS*.

Le codage des adjectifs composés est encore plus complexe que celui des noms, puisqu'à la flexion en genre vient s'ajouter la mise au pluriel. Par ailleurs, certaines unités intègrent des déterminants ou des pronoms qui développent leur propre paradigme (*content de SOI, de MOI, de TOI, de LUI*, etc.). On retrouve les trois types de flexion précédemment identifiés : absence de flexion, flexion terminale et flexion interne :

- absence de flexion : les pluriels obligatoires ne donnent pas lieu à un code flexionnel. Cela ne concerne que quelques mots : *bons amis, comme frère et sœur, comme larrons en foire, en boules de loto, mari et femme, sonnantes et trébuchantes* ;
- flexion terminale : de nombreux adjectifs composés se fléchissent de la même manière que des mots simples : invariabilité (féminin et pluriel identiques au masculin singulier : à *but lucratif*) ; flexion terminale (*à moitié vide, -es ; semi-public, -cs, -que, -ques*) ; variantes (*anti-grève, pl. anti-grève ou anti-grèveS ; semi-nasal, pl. semi-nasaLS ou semi-nasaUX*) ;
- flexion interne : la flexion interne peut concerner n'importe quel élément : ÂPRE *au gain* (flexion sur le 1er élément) ; *plus* MORT *que* VIF (flexion sur le 2e et 3e éléments) ; NOURRI, LOGÉ, BLANCHI NÉCESSAIRE (flexion sur tous les éléments), etc. Par ailleurs, plusieurs cas de flexion peuvent se présenter :

- un même élément peut donner lieu à deux flexions distinctes : *frais émoulu* ➢ *fém. FRAIS émoulue ou FRAÎCHE émoulue*,
- un même élément peut être fléchi ou non fléchi : *taillé en pointe* ➢ fém. sing. *tailléE en pointe,* masc. plur. tailléS en pointe ou en pointeS, fém. plur. tailléES en pointe ou en pointeS ;

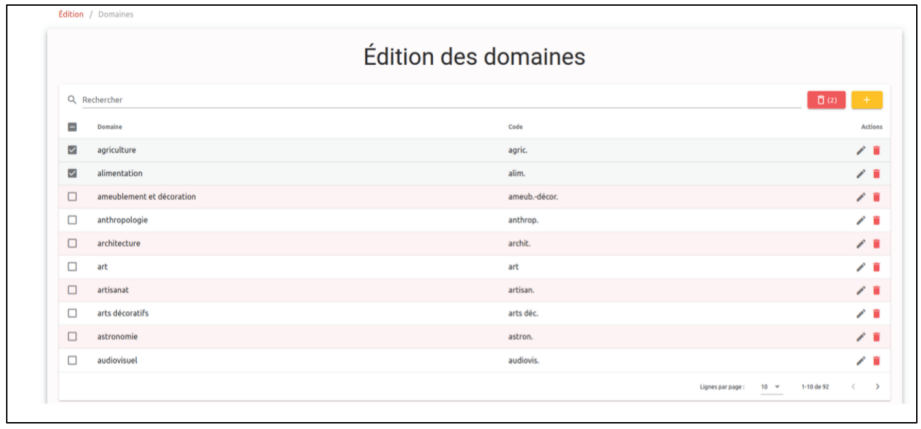
– flexion interne et terminale : une même unité peut cumuler les deux types de flexion : *en épingle à cheveux* ➢ pl. *en ÉPINGLE à cheveux* ou *en ÉPINGLES à cheveux*.

3.4. Implémentation des domaines et des verbes pronominaux

Parmi les dernières fonctionnalités que nous avons ajoutées à Morfetik : l'implémentation des domaines et des formes pronominales.

À partir de l'interface éditeur, il est maintenant possible d'intégrer à la description d'une entrée son domaine, à l'aide d'une liste proposée :

Figure 7
Édition des domaines (image prise sur le site Morfetik)



Nous ne codons les indications de domaine que pour les termes scientifiques ou techniques. Il ne s'agit pas ici d'établir un système conceptuel hiérarchisé, comme celui qu'utilisent les bibliothèques pour le catalogage (par exemple la classification Dewey). Les indications de domaine ont pour fonction essentielle de mieux décrire le lexique (parmi d'autres descripteurs) et non de rendre compte de réalités extra-linguistiques.

Pour l'utilisateur de la ressource, le codage des domaines permet d'identifier automatiquement le vocabulaire relatif à thème donné : il suffit d'extraire toutes les unités lexicales associées à un domaine spécifique. Par exemple, une requête portant sur les termes de la physique dans Morfetik donnera instantanément la réponse suivante : *nombre atomique, neutrino muonique, mesure magnétique, moment angulaire, intelligence fluide, dilatation cubique, échange adiabatique*, *résistance mécanique*, etc. Il devient alors possible d'envisager, à partir de l'ensemble des données enregistrées dans les dictionnaires, la constitution automatique de lexiques spécialisés ou, plus généralement, relatifs à un domaine d'activité particulier.

Enfin, les formes pronominales ont également été implémentées. Au niveau de l'interface d'édition, les verbes sont donc maintenant répartis dans trois tables différentes : les verbes n'ayant pas d'emploi pronominal, les verbes pouvant avoir un usage pronominal et les verbes ayant un usage uniquement pronominal :

Figure 8

Usage pronominal du verbe aimer (image prise sur le site Morfetik)

aimer	Verbe (vrb)		Ind-pr	1		S	

Infinitif	Lemme	Notes
aimer	aimer	

	Indicatif présent		Indicatif imparfait		Passé simple
J' (Je)	(m') aime	J' (Je)	(m') aimais	J' (Je)	(m') aimai
Tu	(t') aimes	Tu	(t') aimais	Tu	(t') aimas Rare
Il / Elle / On	(s') aime	Il / Elle / On	(s') aimait	Il / Elle / On	(s') aima
Nous	(nous) aimons	Nous	(nous) aimions	Nous	(nous) aimâmes
Vous	(vous) aimez	Vous	(vous) aimiez	Vous	(vous) aimâtes Rare

3.6. Matrices morphologiques

Enfin, Morfetik propose des matrices morphologiques, alors que d'autres ressources se contentent de décrire les différentes formes liées à un lemme. Ces matrices sont particulièrement utiles car elles permettent d'étendre la couverture des dictionnaires de manière dynamique, notamment pour les parties du discours lexicales qui constituent des classes ouvertes. Sur corpus, ces matrices permettent une reconnaissance dynamique de formes inconnues, sans avoir à décrire explicitement les formes effectives. Cela est possible grâce aux informations contenues dans les matrices morphologiques, qui capturent les variations morphologiques d'un mot ou d'une forme linguistique.

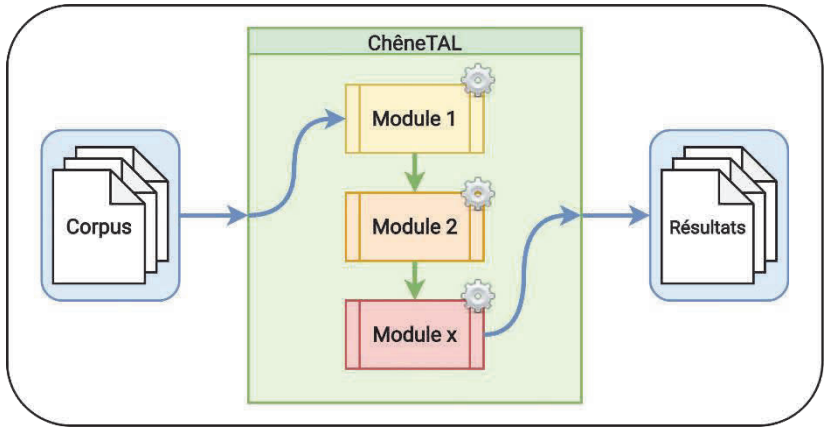
4. Pour ne pas conclure

Dans le cadre d'applications centrées sur les ressources lexicales, l'enjeu est la mise en œuvre de ressources linguistiquement fondées, rigoureuses et actualisables. Le schéma traditionnel d'enrichissement de textes bruts suppose un ensemble de ressources linguistiques de bonne qualité pour l'analyse de la langue naturelle. La plus évidente de ces ressources est un ensemble de formes fléchies, associées à un certain nombre d'informations morphologiques, syntaxiques ou sémantiques, comme Morfetik.

Dans la continuité de ce travail, nous avons développé une plateforme de TAL, nommée ChêneTal, dans laquelle Morfetik est intégrée. Bien souvent, la manipulation des outils TAL est relativement complexe, et peu d'outils conviviaux ont été développés et peuvent être mis entre les mains de non-spécialistes. La plateforme ChêneTal a ainsi été conçue pour permettre la mise en place de chaînes hétérogènes de TAL en intégrant des logiciels existants en gestion et manipulation de corpus (Néoveille, SDMC) avec des modèles plus récents d'Intelligence Artificielle et en proposant une interface simple et intuitive pour les chercheurs de la communauté en TAL et pour ceux en Linguistique/Sciences Humaines et Sociales non spécialistes en informatique.

Cette plateforme intègre divers outils, dont Morfetik, qui offrent à l'utilisateur des possibilités de traiter des corpus et des procédures de recherche :

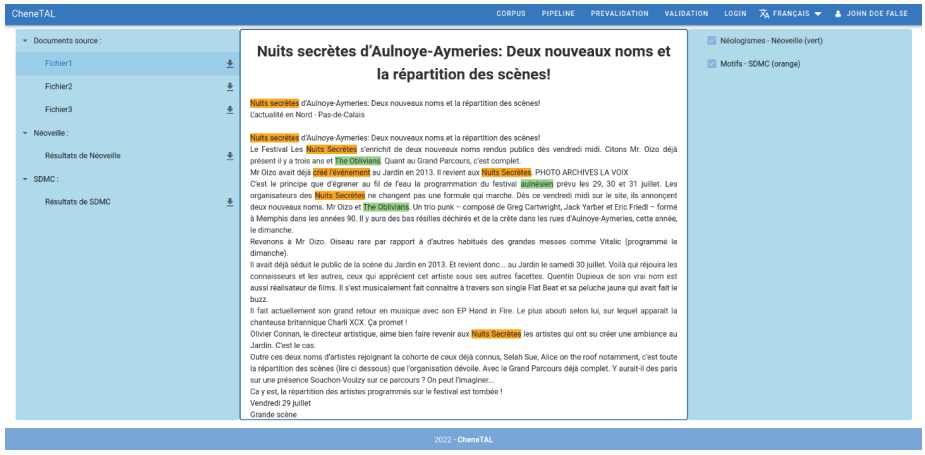
Figure 9
Schéma simplifié du fonctionnement de ChêneTAL



L'utilisateur de ChêneTAL choisit un corpus (fourni par la plateforme ou importé par l'utilisateur) et définit une chaîne de traitement, c'est-à-dire une suite ordonnée de modules de traitement paramétrés par l'utilisateur. Ces modules sont des outils TAL développés au sein de l'équipe RCLN du LIPN[11] comme Néoveille (outil de détection et de suivi des néologismes dans la presse en ligne) et SDMC (outil de fouille de données pour l'analyse de corpus et l'extraction de motifs). Une fois que le traitement est complété, ChêneTAL affiche l'ensemble des résultats retournés par les modules de traitement dans une interface intuitive :

Figure 10

Interface de visualisation des résultats de ChêneTAL (image prise sur le site ChêneTAL)



Morfetik est donc une plateforme en constante évolution qui vise à se développer progressivement pour améliorer le traitement des données textuelles.

Références citées

Buvet, P.-A., Cartier, E., Issac, F., Mathieu-Colas, M., Mejri, S. & Madiouni, Y. (2009). Morfetik, ressource lexicale pour le TAL. *16ème Conférence sur le Traitement Automatique des Langues Naturelles, Actes du colloque TALN 2009, Juin 2009, Senlis (France)*, 1–10.

[11] Laboratoire d'Informatique de Paris Nord, LIPN, UMR CNRS 7030, Université Sorbonne Paris Nord.

Cartier, E. (2011). Néologie et description linguistique pour le TAL. *Langages 3*(183), 105–117.

Cartier, E. (2019). Néoveille, plateforme de repérage et de suivi des néologismes en corpus dynamique. *Neologica : revue internationale de la néologie 13*, 23–54.

Cartier, E. & Sablayrolles, J.-F. (2009). Néologismes, dictionnaires et informatique. *Cahiers de Lexicologie 2008-2*(93), 175–192.

Clément, L., Lang, B., & Sagot, B. (2004). Morphology based automatic acquisition of large-coverage lexica. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, 1841–1844.

Courtois, B. (1990). Un système de dictionnaires électroniques pour les mots simples du français. *Langue française 87*, 11–22.

Grezka, A. (2017). Morfetik, un dictionnaire morphologique : illustration avec le lexique de la perception. Dans E. Biardzka, M. Dańko, G. Komur-Thilloy & F. Marsac (éds), *La Perception en langue et en discours, Echo des études romanes* (89–99). Jihočeská univerzita.

Grezka, A. (2020). Variabilité et traitement automatique des langues. *Lingvisticae Investigationes 43*(2), 280–299.

Grezka, A., Cartier, E. & Mathieu-Colas, M. (2015). Dictionnaires morphologiques du français contemporain : présentation de Morfetik, éléments d'un modèle pour le TAL, *Actes du colloque TALN 2015, 22–25 juin 2015, Université de Caen Basse-Normandie, Caen (France)*, 466–472.

Gross, G. (1996). *Les expressions figées en français noms composés et autres locutions*. Ophrys.

Hathout, N., Sajous, F. & Calderone, B. (2014). GLÀFF, a Large Versatile French Lexicon. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 1007–1012.

Mathieu-Colas, M. (1990). Orthographe et informatique : établissement d'un dictionnaire électronique des variantes graphiques. *Langue française 87*, 104–111.

Mathieu-Colas, M. (2009). Morfetik : une ressource lexicale pour le TAL. *Cahiers de Lexicologie 94*, 137–146.

Mathieu-Colas, M. (2014). Flexion des noms composés : principes de codage. Dans Z. Gavriilidou & A. Revithiadou (éds), *Studies dedicated to Professor Emeritus Anna Anastasiadis-Symeonidis of the Aristotle University of Thessaloniki*, 196–210.

Sagot, B. (2010). The *Lefff*, a freely available and large-coverage morphological and syntactic lexicon for French. *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, 2744–2751.

Sajous, F., Hathout, N. & Calderone, B. (2013). GLÀFF, un Gros Lexique À tout Faire du Français. *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN 2013)*, 285–298.

Sajous, F., Hathout, N. & Calderone, B. (2014). Ne jetons pas le Wiktionnaire avec l'oripeau du Web ! Études et réalisations fondées sur le dictionnaire collaboratif. *Actes du 4e Congrès Mondial de Linguistique Française (CMLF 2014)*, 663–680.