

# Morfetik : Une ressource lexicale morphologique extensible et modulaire pour le français

Jaime Arias\*, Othman Boudarga\*, Aude Grezka\*

\*CNRS, Université Sorbonne Paris Nord, LIPN, F-93430 Villetaneuse, France  
{arias,boudarga,grezka}@lipn.univ-paris13.fr

**Résumé.** Les ressources lexicales morphologiques, décrivant la structure interne des mots et leurs formes fléchies, sont essentielles pour le traitement automatique des langues (TAL) et la linguistique computationnelle.

Nous présentons MORFETIK, une ressource lexicale open-source complète pour le français, capable de générer et d’identifier automatiquement toutes les formes fléchies des mots (noms, verbes, adjectifs, locutions, etc.). Il offre une couverture large du lexique contemporain et spécialisé, une architecture extensible et modulaire, et une intégration aisée avec des ressources externes.

De même, nous illustrons son utilisation à travers deux études de cas et détaillons son architecture, montrant comment sa modularité et son interopérabilité facilitent l’analyse de corpus, et le développement d’applications TAL.

## 1 Introduction

Une *ressource lexicale* est une base de données linguistique structurée qui rassemble des informations sur les mots d’une langue, leurs formes, leurs significations et leurs relations. Elle constitue un élément essentiel pour la recherche en traitement automatique des langues (TAL), en linguistique computationnelle et en technologies du langage. Selon leur conception et leurs objectifs, les ressources lexicales peuvent décrire différents aspects du lexique : la *morphologie* (formes et flexions des mots), la *syntaxe* (régimes et constructions), la *sémantique* (sens et relations lexicales) ou encore la *pragmatique*.

Dans cet article, nous présentons MORFETIK, une *ressource lexicale morphologique* complète et modulaire pour le français, conçue pour générer, structurer et exploiter automatiquement les formes fléchies du lexique (noms, adjectifs, déterminants, pronoms, verbes, adverbes, prépositions, conjonctions, interjections, locutions, etc.). Elle permet d’obtenir, pour n’importe quel mot français, l’ensemble de ses formes (pluriel des noms, féminin et pluriel des adjectifs, formes conjuguées des verbes, etc.), ou bien, réciproquement, d’identifier le mot (la forme de base, le “*lemme*”) correspondant à n’importe quelle forme fléchie.

MORFETIK est une ressource clé pour l’analyse, la recherche et le développement d’applications en TAL qui offre :

- une couverture lexicale très large du français contemporain et spécialisé (médecine, minéralogie, etc.), plus de 240.000 lemmes ;