Introduction

Maria Grossmann, Fabio Montermini, Diana Passino, Diego Pescarini et
Florence Villoing

Édition électronique

URL : <https://journals.openedition.org/corpus/6823>

DOI : 10.4000/corpus.6823

ISSN : 1765-3126

Éditeur

Bases ; corpus et langage - UMR 6039

Référence électronique

Maria Grossmann, Fabio Montermini, Diana Passino, Diego Pescarini et Florence Villoing,
« Introduction », *Corpus* [En ligne], 23 | 2022, mis en ligne le 01 mars 2022, consulté le 05 mars 2022.
URL : <http://journals.openedition.org/corpus/6823> ; DOI : <https://doi.org/10.4000/corpus.6823>

Ce document a été généré automatiquement le 5 mars 2022.

© Tous droits réservés

Introduction

Maria Grossmann, Fabio Montermini, Diana Passino, Diego Pescarini et
Florence Villoing

Nous tenons à remercier chaleureusement tous les collègues qui ont bien voulu assurer les relectures des articles de ce numéro pour leur contribution précieuse.

À l'instar d'autres disciplines linguistiques, la morphologie a connu, au cours des deux dernières décennies, une véritable 'révolution empirique'. Puisque les mots, leur structure et les relations qu'ils entretiennent constituent l'objet central d'étude des morphologues, la collecte et l'analyse de données, notamment lexicales, ont toujours représenté un enjeu important pour cette discipline. Cependant, si par le passé, elle a trouvé dans les ressources que l'on peut qualifier de 'traditionnelles' (dictionnaires, corpus littéraires ou journalistiques, etc.) d'importantes sources pour des études empiriques, y compris à grande échelle, elle connaît aujourd'hui une évolution, à la fois quantitative et qualitative, dans la collecte et la manipulation des données qui modifie profondément la nature des questions qui sont adressées, les méthodes employées et les explications proposées. Dans certains cas, cette évolution a conduit à de véritables changements de paradigme et à une redéfinition des unités impliquées dans l'analyse. Plusieurs facteurs ont impulsé cette dynamique. D'une part, l'accès à des quantités de données inimaginables, jusqu'à il y a quelques décennies, est devenu de plus en plus simple et rapide, notamment via la Toile, et ce en parallèle à un développement notable des outils d'extraction, d'annotation et d'analyse. D'autre part, l'interaction avec des disciplines aussi diverses que l'informatique, les mathématiques, l'étude de l'acquisition du langage, la psycholinguistique, les neurosciences (pour ne citer que les principales), a conduit à un changement profond des méthodes et des concepts mobilisés. Ce mouvement a touché à la fois l'étude de la morphologie flexionnelle et de la morphologie dérivationnelle (qui intègre autant la dérivation affixale, que la conversion et la composition), ainsi que la plupart des sous-domaines et approches théoriques.

- 1 L'objectif principal de ce numéro de *Corpus* est de tracer un bilan de l'état actuel des études en morphologie, à la lumière de l'évolution décrite ci-dessus. La variété des travaux qui y sont accueillis témoigne, d'une part, de la vitalité de ce champ d'études, et d'autre part du fait que le mouvement que nous avons qualifié de révolution

empirique touche, sinon l'ensemble, du moins la majorité des domaines d'étude et des approches. Les articles que nous avons rassemblés couvrent, en effet, un large spectre qui va de la constitution de ressources pour l'analyse morphologique des langues, aux approches quantitatives et statistiques, en passant par l'analyse diachronique et l'étude des langues minoritaires. Ils témoignent en outre du fait que la réalisation de bases de données riches et exploitables est désormais courante, non seulement pour les langues de large diffusion et bénéficiant d'une longue tradition d'études (français, italien, portugais, roumain, russe), mais également pour des langues minoritaires et jusqu'à présent peu dotées (corse, variétés dialectales du finnois, variétés walser du Nord de l'Italie).

- 2 Le numéro s'ouvre avec la contribution d'Alina Villalva et Rafael Dias Minussi (*Description and analysis of a Portuguese blend corpus*), consacrée à l'analyse des caractéristiques des mots-valises (« *blends* », dans la littérature anglo-saxonne) en portugais. Leur recherche se fonde sur un corpus construit *ad hoc*, d'environ 300 mots-valises tirés de diverses sources relevant des différentes variétés diatopiques du portugais (européen, brésilien, angolais et mozambicain). La description des caractéristiques saillantes des différents types de mots-valises et de leurs constituants est accompagnée de l'illustration des résultats de deux tests expérimentaux ayant pour but d'investiguer l'interprétation et le traitement de ces lexèmes par des étudiants universitaires portugais et brésiliens.
- 3 Suivent ensuite trois articles qui abordent, de manières différentes, des questions de diachronie. Le premier, celui de Silvia Micheli (*CODIT. A new resource for the study of Italian from a diachronic perspective: Design and applications in the morphological field*), exploite un nouveau corpus diachronique de la langue italienne : CODIT (*Corpus Diaconico dell'Italiano*). CODIT est un corpus de 29 millions d'occurrences qui est structuré en cinq sous-corpus ; chaque sous-corpus est constitué de textes appartenant au même stade diachronique. Après une présentation détaillée du corpus, de ses objectifs et de la méthodologie à la base de sa construction, l'article est consacré à deux études de cas : la perte progressive de la valeur évaluative du préfixe *sopra-* et la diffusion du suffixe *-oide* qui, à l'origine, était réservé à la terminologie scientifique. L'objet principal du deuxième article, *Towards a diachronic analysis of Romance morphology through Google n-grams*, de Jan Radimský, est l'utilisation du corpus Google n-grams pour l'analyse diachronique de la morphologie du français et de l'italien. L'auteur présente, en particulier, les premiers résultats quantitatifs de l'extraction des données pour les deux langues en question (comprenant les métadonnées associées), en discutant les problèmes liés à l'extraction d'unités dans un corpus de ces dimensions et caractéristiques, et en se focalisant en particulier sur les mots morphologiquement complexes, dont une partie apparaît comme des suites de deux mots graphiques distincts. Et enfin, le troisième article d'Adrian Chircu, *Les adjectifs-adverbes hérités du latin en ancien roumain. Le XVI^e siècle*, étudie l'inventaire et les caractéristiques des adjectifs utilisés en tant qu'adverbe dans les premiers témoignages écrits du roumain. L'étude se fonde sur des données tirées du *Corpusul limbii române vechi* (Corpus de l'ancien roumain), élaboré par l'Institut de linguistique « Iorgu Iordan – Alexandru Rosetti » de l'Académie roumaine, constitué de textes allant de 1521 à 1780. L'auteur montre que l'usage généralisé de ce procédé morphologique, qui distingue le roumain d'autres langues romanes dans lesquelles la transcatégorisation des adjectifs est présente seulement marginalement, est bien attesté déjà en ancien roumain.

- 4 Les trois articles qui suivent s'intéressent à la place des corpus dans l'étude des langues minoritaires, que ce soit du point de vue de la constitution de ressources en vue de l'étude de leur morphologie ou de l'exploitation de ressources existantes. Livio Gaeta, Marco Angster, Raffaele Cioffi et Marco Bellante, dans leur article *Corpus linguistics for low-density varieties. Minority languages and corpus-based morphological investigations*, présentent des projets d'acquisition et d'analyse de données de variétés en danger, parlées par les communautés allemandes walser du Piémont et de la Vallée d'Aoste, à travers la linguistique de corpus. Ils illustrent d'abord les enjeux, les difficultés et l'apport de l'application de la linguistique de corpus aux variétés en danger, caractérisées typiquement par une granularité importante et des corpus de petite taille, et discutent des solutions originales mises en place pour la construction de la plateforme, qui à travers la création d'outils raffinés mais pratiques, permet d'effectuer des analyses morphologiques, syntaxiques et lexicales basées sur corpus. Des études de cas sont exposées successivement, notamment concernant les tendances observées dans la distribution des formes des préfixes verbaux et leur représentation, l'introduction du niveau 'mot phonologique' pour gérer la variation morphophonologique des séquences verbe-clitique et les schémas de variation des participes passés, qui montrent des développements intéressants vis-à-vis du comportement flexionnel des classes verbales. Jean Léo Léonard, dans son article *Parangons inférentiels dans l'atlas linguistique du finnois de Lauri Kettunen. Une approche réalisationnelle en morphologie flexionnelle d'un corpus dialectal fennique*, explore un échantillon de données issu de l'atlas des dialectes finnois de Lauri Kettunen (1930, 1940). L'analyse proposée par Léonard démontre qu'un outil de recherche qui a été conçu selon une architecture néogrammairienne est un thésaurus de données exploitables pour l'étude de la morphologie flexionnelle dans le cadre de la Morphologie des Fonctions Paradigmatiques (Stump 2001, 2015). Les données analysées par Léonard montrent que l'évolution et la diversification des dialectes finnois est contrainte et prédictible selon des mécanismes syllabiques donnant lieu à la formation de nouveaux thèmes flexionnels.
- 5 L'article de Stella Retali-Medori et Laurent Kevers, *La morphologie dans la Banque de Données Langue Corse : bilan et perspectives*, après avoir parcouru les phases de la création et du développement de la Banque de Données Langue Corse (BDLC) au sein du *Nouvel Atlas Linguistique de la Corse* (NALC), se concentre sur le traitement de la morphologie dans les différentes versions de la BDLC. Une réflexion est ensuite entamée sur les évolutions possibles du module morphologique de l'actuelle version de la base de données. Le module envisagé permettrait d'accroître le potentiel d'exploitation de la base à travers, entre autres, un traitement cartographique des données, particulièrement intéressantes du point de vue de la morphologie flexionnelle ou dérivationnelle. Citons, à titre d'exemple, les alternances de genre (qui révèlent les restructurations opérées depuis le neutre latin, dans les parlers corses) et des combinaisons de racines et d'affixes particulières, telles que celles contenant des chaînes de suffixes. Dans le cadre de l'évolution du module morphologique, les applications du TAL au corse, explorées dans la deuxième partie de l'article, revêtent une importance particulière dans la mesure où, comme le discutent en détail les auteurs, cela permet aussi bien l'exploitation de la base de données, que le développement d'outils facilitant leur saisie et leur analyse.

- 6 Le numéro se termine avec deux articles présentant des approches statistiques sur des ensembles de données extensives pour l'étude de la morphologie dérivationnelle, du russe et du français, respectivement. Dans son article *Statistical modelization of suffixal rivalry in Russian: adjectival formations in -sk- and -n-*, Natalia Bobkova présente les résultats d'une investigation de la concurrence suffixale dans la construction d'adjectifs dénominaux en russe – en particulier concernant l'alternance entre les deux suffixes les plus productifs, *-sk-* et *-n-* – sur la base de données de haute et de basse fréquence tirées du *Corpus national de la langue russe*. L'analyse permet à l'auteure d'identifier les propriétés des noms de base qui apparaissent comme les plus significativement corrélées au choix du suffixe, et de vérifier l'intérêt d'étudier les mots de basse fréquence en tant qu'indicateurs de l'usage créatif que les locuteurs font des procédés morphologiques. L'article de Marine Wauquier, *Apports de la sémantique distributionnelle pour la morphologie dérivationnelle*, quant à lui, dresse un état des lieux de l'utilisation de la sémantique distributionnelle en morphologie dérivationnelle. L'auteure montre en quoi l'évolution récente des outils permettant de constituer de larges corpus et bases de données, ainsi que le développement d'outils statistiques, ont rendu possible une approche extensive et quantitative de la morphologie. La sémantique distributionnelle, en offrant une approche quantitative et statistique du sens des mots morphologiquement construits à partir de larges corpus, participe de ce mouvement et permet d'appréhender, dans une perspective quantitative nouvelle, des problématiques anciennes comme inédites. L'article questionne ainsi les possibilités d'une telle approche et signale, en outre, les précautions qu'elle nécessite.
-

BIBLIOGRAPHIE

- Kettunen L. (1930). *Suomen murteet II. Murrealueet*. Helsinki : Suomalaisen Kirjallisuuden Seuran Toimituksia.
- Kettunen L. (1940). *Suomen murteet III. A. Murrekartasto ja B. Selityksiä murrekartastoon*. Helsinki : Suomalaisen Kirjallisuuden Seuran Toimituksia.
- Stump G.T. (2001). *Inflectional Morphology. A Theory of Paradigm Structure*. Cambridge : Cambridge University Press.
- Stump G.T. (2015). *Inflectional Paradigms. Content and Form at the Syntax-Morphology Interface*. Cambridge : Cambridge University Press.