

Les grands corpus oraux, pour quoi faire ?

Depuis plus d'un siècle, les linguistes collectent des enregistrements sonores (des corpus oraux) afin de décrire les langues dans toute leur variété et de réaliser des applications diverses, de l'enseignement jusqu'au traitement automatique des langues. Depuis une vingtaine d'années, les études sur les corpus de langues parlées ont complètement renouvelé les sciences du langage, tant sur le plan descriptif que théorique et méthodologique. À l'ère du numérique, on dispose d'outils informatiques permettant la classification des données et leur accès aisé, offrant du même coup la possibilité de sauver ce patrimoine et de le valoriser en transformant les documents originaux en véritables ressources linguistiques. Si la plupart des grandes et moins grandes langues d'Europe disposent de grands corpus oraux numérisés et accessibles en ligne (souvent gratuitement), le français ne dispose jusqu'à ce jour que de petits corpus de conceptions hétérogènes et peu accessibles. La France, qui était en avance pour la mise au point des corpus de langue écrite (en particulier pour le *Trésor de la langue française*) a pris du retard dans la constitution des corpus de langue parlée. Depuis 2004, le programme « corpus de la parole » du ministère de la Culture et de la Communication (Délégation générale à la langue française et aux langues de France, avec le soutien du **plan national de numérisation**) favorise la conservation, la mise à disposition et la valorisation des corpus oraux existants du français et des 75 langues de France (<http://corpusdelap parole.culture.fr>). Toutefois, un grand corpus de français parlé reste à construire.

Qu'est-ce qu'un corpus oral ?

Un grand corpus de langue parlée est constitué d'échantillons enregistrés d'une langue dans toutes ses variétés (régions, genres, locuteurs) et de leurs transcriptions. L'échantillonnage doit être fait selon des critères réfléchis et en vue d'utilisations diverses. La transcription peut répondre à divers critères de finesse, mais doit être effectuée par des professionnels. Le corpus lui-même est accompagné de divers outils informatiques : alignement texte/son, concordanciers nécessaires à l'analyse lexicale, analyses syntaxiques et prosodiques, logiciels d'exploitation. Une fois établi, il doit



être géré par un organisme responsable de son exploitation et de sa conservation. La constitution d'un tel corpus est une entreprise coûteuse : les spécialistes en estiment le prix à un euro par mot... Le jeu en vaut-il vraiment la chandelle ? Sans aucun doute.

À quoi sert un corpus oral ?

Dans le domaine de la recherche linguistique, les grands corpus oraux servent en premier lieu comme base de données pour la connaissance de la langue et pour son enseignement. Ils permettent la description des usages réels dans toute leur diversité et complexité et l'étude du devenir d'une langue dans toutes ses dimensions (phonologique et prosodique, lexicale et syntaxique, sémantique et pragmatique). C'est aussi un instrument indispensable pour la comparaison des langues et l'étude de la variation sous toutes ses formes (par exemple : français en France/hors de France, régions/capitale, enfants/adultes, usages « normaux »/« perturbés »).

Les grands corpus oraux servent aussi comme base de données pour les industries du langage : reconnaissance et synthèse de la parole, traitement automatique des langues (applications aux dialogues homme/ machine, services vocaux d'information, traductions automatiques). Ils sont indispensables pour la diffusion de la langue à un niveau international.

La dimension utile des corpus varie selon l'approche. Mais si l'on veut étudier le lexique, ou bien des corrélations entre le langage et d'autres phénomènes, il faut un ensemble vaste et diversifié. Les grands corpus de langue parlée collectés aujourd'hui dans le monde comptent dix millions de mots transcrits (1 000 heures d'enregistrement). C'est le cas pour le *British National Corpus (BNC)*, le *Corpus de Referencia del Español Actual (CREA)*, le *Corpus Gesproken*

Nederlands (CGN). Les corpus actuels de français parlé ne dépassent pas 2 millions de mots : les recherches lexicales y sont donc réduites et les données statistiques peu fiables. Cela ne permet pas non plus de comparer efficacement le français à d'autres langues. Le français risque de la sorte d'être absent des études comparatives et typologiques et des diverses applications, tant sur le plan des traductions qu'en ingénierie linguistique.

Ainsi, la constitution et la sauvegarde d'une base de données orales variée, fiable et accessible sont des enjeux de première importance, tant sur le plan de la recherche que pour le développement d'une politique culturelle qui reconnaisse les faits de langue dans toute leur variété comme éléments du patrimoine immatériel.

Dominique Willems

Université de Gand

Académie royale de Belgique

The uses for major spoken corpora

The study of spoken language corpora over the last few years has breathed new life into language sciences. In France, the "Corpus de la parole" programme run by the Ministry of Culture and Communication (DGLFLF) promotes the conservation and availability of existing spoken corpora for French and the 75 languages of France. However, a major spoken corpus, accessible online, does not yet exist for French, as it does for English, Spanish and Dutch. A spoken corpus brings together recorded samples of a language's variations and their transcription, and provides a database to promote the understanding and teaching of the language. It is used to describe actual usage of a language, the study of its evolution and comparisons. It is also useful for language industries and plays a vital role in international dissemination.

Wozu dienen mündliche Sprachkorpora ?

In den letzten Jahren haben die Studien zu den Korpora gesprochener Sprachen die Sprachwissenschaften revolutioniert. In Frankreich fördert das vom Ministerium für Kultur und Kommunikation (DGLFLF) entwickelte Programm corpus de la parole („Korpus gesprochener Sprache“) die Erhaltung und Bereitstellung der bestehenden mündlichen Textkorpora für Französisch und die 75 Sprachen Frankreichs. Doch während ein solcher Korpus für Englisch, Spanisch oder Niederländisch bereits existiert, muss ein umfassender und online zugänglicher mündlicher Textkorpus für das Französische erst erarbeitet werden. Ein mündlicher Sprachkorpus vereint Tonaufzeichnungen aller Sprachvariationen und ihre Transkription. Er bildet eine Datenbasis für Spracherwerb und -unterricht und ermöglicht die Beschreibung der realen Anwendungen in einer Sprache, Studien zur Sprachentwicklung und den kontrastiven Sprachvergleichen. Zudem dient er den kommerziellen Sprachanbietern für den internationalen Vertrieb.