Le Projet TELOC : construction d'une base textuelle occitane

Myriam Bras
CLLE-ERSS (UMR 5263 CNRS) & Université de Toulouse II le Mirail
myriam.bras@univ-tlse2.fr

De plus en plus de langues ont leurs bases textuelles (cf. Frantext, *The British National Corpus, El Corpus Textual Informatitzat de la Llengua Catalana, XX. mendeko euskararen corpus estatistikoa, Base de Datos Sintácticos del español actual…* pour ne citer que quelques exemples). Ce sont des ressources indispensables à toute entreprise lexicographique, et à toute description scientifique de la langue dans ses dimensions lexicale, morphologique, syntaxique, sémantique et discursive.

Le projet TELOC (Textes En Langue OCcitane) a pour ambition de doter l'occitan d'une telle ressource. Il vise la constitution d'une base de données textuelles regroupant des œuvres écrites de tous genres (littérature, théâtre, conte, textes techniques, journalistiques,…). Il est mené par l'ERSS, qui allie des compétences en linguistique occitane, en linguistique de corpus et en traitement automatique des langues.

La base textuelle de TELOC sera consacrée à l'occitan moderne et contemporain écrit. Le corpus à réunir est immense : on estime à plusieurs milliers le nombre d'œuvres produites en occitan depuis le 16^{e} siècle. Pour les œuvres antérieures, la totalité du corpus est en passe d'être rassemblée dans la base textuelle du projet *Concordance de l'Occitan Médiéval,* sous la direction du Professeur Ricketts.

La première étape du projet est la construction d'une base expérimentale de taille modeste comportant une vingtaine d'œuvres (environ un million de mots). Il s'agit de regrouper des œuvres contemporaines, donc déjà sous format numérique, et de les coder en XML selon une norme internationale (*Text Encoding Initiative* P5). Cette phase d'expérimentation est menée en partenariat avec l'ATILF à Nancy sur le modèle d'une base textuelle de type Frantext. La base sera accessible au public dans le cadre du Centre national de ressources textuelles et lexicales. TELOC bénéficiera ainsi d'une mutualisation des techniques et des outils au sein du CNRTL, et en particulier des outils d'interrogation de la base textuelle. Dans cette version expérimentale, la base pourra être exploitée par des requêtes simples : extraction de contextes contenant un mot, une partie ou une séquence de mots, recherche de co-occurrences, calcul de fréquences de mots.

Les étapes suivantes viseront un accroissement progressif et significatif de la base : il y a suffisamment de matière pour envisager de passer, à moyen ou long terme, à plusieurs centaines de millions de mots. L'augmentation du volume des données s'accompagnera d'une structuration de la base : classement par genre et par domaine, par type de support, par époque et par date, par dialecte et variante, par type de graphie adoptée par l'auteur. On veillera à ce que la base respecte, à terme, les règles de constitution d'un véritable corpus : représentativité équilibrée des genres, des domaines, des dialectes, des types de support, présence de textes oraux. Une base de textes oraux, avec laquelle nous envisageons d'interfacer TELOC, est en cours de constitution dans le cadre du THESOC (THEsaurus OCcitan mené par l'UMR 6039 à Nice et l'ERSS).

Toutefois, dans les étapes qui suivront immédiatement la phase expérimentale, on s'attachera à fixer certains paramètres : on pourra par exemple commencer par réunir et coder des textes en languedocien de l'époque contemporaine et ne considérer pour commencer que des écrits fictionnels sur support de type livre. Sur la base élargie ainsi constituée, on pourra passer à une phase de traitement linguistique qui permettra d'enrichir la base avec des informations morpho-syntaxiques pour en faire une base catégorisée. Le travail d'étiquetage morpho-syntaxique du corpus permettra de lemmatiser la base et de l'exploiter avec des requêtes plus complexes : par exemple, chercher toutes les formes fléchies d'un verbe, d'un nom ou d'un adjectif ; chercher les dérivés d'un mot ; sélectionner des séquences de phrases conjuguées uniquement avec certains temps verbaux. Le moteur de recherche devra intégrer le haut degré de variation spécifique à l'occitan : variation graphique, mais aussi flexionnelle et lexicale.

Outre les laboratoires déjà cités, TELOC a pour partenaires l'Institut d'études occitanes (IEO-IDECO) et le Centre de ressources occitanes et méridionales. Des collaborations sont souhaitées avec les artisans de projets proches comme la bibliothèque virtuelle du CIEL d'OC et de l'université de Provence, le dictionnaire informatisé du GIDILOC et, plus largement, avec tous les chercheurs motivés par un tel projet. Car au-delà des ses utilisations pour la linguistique occitane, la linguistique romane comparatiste, la lexicographie, TELOC sera précieux pour l'enseignement de la langue, pour des études littéraires, ethnologiques, historiques, pour la sauvegarde du patrimoine écrit et la mise à la disposition du public des écrits occitans.

In *Langues et cité : bulletin de l'observatoire des pratiques linguistiques*, 8, Décembre 2006, p.9.