

OBJECTIFS DU CORPUS

Modéliser le changement : les voies du français

France Martineau: directrice

Le corpus *Voies du français* est structuré de façon dialectale, sociale et historique et constituera, lorsqu'il sera achevé, une base de plus de 2,5 millions de mots, base assez large pour tirer profit d'analyses statistiques sur différentes variables. Le corpus couvre quatre périodes historiques:

- I. Ancien français
- II. Moyen français
- III. Français du XVI^e siècle
- IV. Français de la Nouvelle-France (XVII^e et XVIII^e siècles)

Les textes sont d'abord tous soumis à un balisage de base, de façon à uniformiser les formats (indication du titre, auteur, date, origine, type de texte, début et fin de texte, etc.). À ce balisage de base, nous ajoutons des métadonnées historiques permettant de préciser le profil social du scripteur (âge, degré d'éducation, lieu de naissance, emploi, conjoint, mobilité, etc.) et le parcours du texte.

Dans le cadre d'une théorie du changement linguistique qui affecte des structures plutôt que des suites de mots, le modèle doit permettre d'analyser de façon automatique des régularités catégorielles et de dévoiler les structures de la langue. C'est pourquoi nous procédons d'abord à une catégorisation grammaticale. Les textes sont alors prêts pour l'étiquetage automatique des phrases en constituants syntaxiques (la structure syntaxique et la fonction des diverses unités : sujet, objet, etc.).

À ce modèle du changement linguistique, nous ajoutons un module *Discours identitaire*. Ce module, parce qu'il découle d'un corpus représentatif de différentes sphères de la société et que nous l'aurons balisé avec des descripteurs historiques fins, permettra de répondre à des questions cruciales sur l'interaction entre l'évolution de certains thèmes qui définissent l'identité et le changement linguistique.