La variation typologique : analyse systématique d'un corpus québécois

Margareta Kastberg Sjöblom

ATST, EA 3187 – Université de Franche-Comté – Besançon – France

Abstract

The purpose with this paper is to explore typological variations and oppositions in a large data base from Quebec, the BDTS Corpus. This normalized and tagged corpus contains a rich variety of texts and oral corpora in French and it represents various genres or typologies. The automatic data-processing and the exploration of a lexico-statistical tool make it possible to observe different linguistic models in the various text types. The different statistical analyses show that the opposition between the various typologies is always present and often even dominating, not only from a lexical or semantic point of view, but even seen from a morphological or syntactic angle. Finally, the measurement of the intertextual distance provides a robust and reliable tool for measuring similarities as well as dissimilarities within large text bases in between different types of text.

Résumé

Le présent exposé propose d'étudier les variations et les oppositions typologiques dans une base de données multigénérique québécoise, la BDTS, en s'appuyant sur un corpus informatisé et lemmatisé, et en exploitant les techniques quantitatives. Ce corpus présente une riche variété de textes et de corpus oraux et se décline en différents genres.

Le recours à des outils de traitement informatiques ou mathématiques particuliers permet de dégager les modèles de fonctionnement (lexical, grammatical, syntaxique, phraséologique, pragmatique) des discours et des textes. L'analyse du corpus en situation montre en effet que le lexique, la morphosyntaxe, la structure et le rythme du récit varient avec les genres. L'opposition entre les différentes typologies est toujours présente et souvent même prépondérante dans les différentes analyses statistiques, notamment lorsque l'on s'intéresse comme ici à la sémantique et aux structures thématiques.

Mots-clés : typologie textuelle – linguistique de corpus – lexicométrie – textométrie – distances intertextuelles

1. Introduction

Cet communication s'intéresse aux variations et aux oppositions génériques dans une base de données multigénérique québécoise, en s'appuyant sur un corpus informatisé et lemmatisé, et en exploitant les techniques quantitatives. Ce corpus présente une riche variété de textes et de corpus oraux et se décline en différents genres.

Le recours à des outils de traitement informatiques ou mathématiques particuliers permet de dégager les modèles de fonctionnement (lexical, grammatical, syntaxique, phraséologique, pragmatique) des discours et des textes pour une meilleure compréhension des genres discursifs et de la textualité. L'analyse des corpus en situation montre en effet que le lexique, la morphosyntaxe, la structure et le rythme du récit varient avec les genres. L'opposition entre les différentes typologies est toujours présente et souvent même prépondérante dans les différentes analyses statistiques, notamment lorsque l'on s'intéresse comme ici à la sémantique et aux structures thématiques.

Afin de compléter les possibilités de recherche déjà offertes par la mise en réseau du corpus par le Secrétariat à la politique linguistique du Québec, et étant donné la riche variation typologique des textes et la relative homogénéité de taille des différents sous-corpus, ce corpus nous a semblé particulièrement intéressant à soumettre à un traitement lexicométrique et textométrique. L'exploration statistique du logiciel lexico-statistique donne la possibilité d'analyses diverses non seulement traditionnelles comme celles sur la richesse lexicale, l'accroissement lexical, la corrélation chronologique etc., mais plusieurs fonctions thématiques sont également disponibles. Le programme permet d'analyser les spécificités lexicales aussi bien d'un point de vue aussi bien endogène qu'exogène, il permet d'extraire les corrélats thématiques et recense aussi tous les termes situés dans l'environnement immédiat d'un mot donné. L'analyse des corrélats sémantiques et thématiques révèle des caractéristiques de chaque typologie présente dans ce corpus et l'analyse factorielle montre que les mêmes orientations des textes se retrouvent aussi bien au niveau lexical et syntaxique qu'au niveau thématique.

L'analyse de la distance entre les différents textes de ce corpus montre également des différences typologiques importantes. Cette étude qui traditionnellement ne concernait que le lexique, s'ouvre désormais à d'autres paramètres tels que

La lemmatisation du corpus et l'accès aux codes grammaticaux permettent aussi d'étudier la morphologie et la syntaxe des différents sous-corpus, une analyse qui met en exergue la division typologique de ces textes. L'analyse du corpus en situation montre en effet que le lexique, la morphosyntaxe, la structure et même le rythme de récit varient avec les genres ou les typologies. L'opposition entre les différentes typologies est toujours présente et souvent même prépondérante dans les différentes analyses statistiques.

2. Le corpus

Parmi les grands projets de recherche sur la langue française au Canada, le secrétariat à la politique linguistique du Québec, grâce au Programme de développement, met en réseau et propose l'exploitation de corpus lexicaux québécois. Ce réseau est constitué de quinze corpus provenant de cinq universités québécoises : l'Université de Laval, l'Université de Montréal, l'Université du Québec à Montréal, l'Université du Québec à Rimouski et l'Université de Sherbrooke.

La base que nous nous proposons d'exploiter ici est un des sous-corpus de *La Banque de Données Textuelles de Sherbrooke*. Ce corpus contient environ deux millions d'occurrences (61 843 formes) tirées de 1054 textes différents, et elle constitue donc un sous-ensemble de la *Banque de Données Textuelles de Sherbrooke* (BDTS), qui contient plus de 16 millions d'occurrences à l'heure actuelle.

Ce corpus est composé de huit sous-ensembles d'environ 250 000 mots chacun et traités selon une norme commune, ce qui rend leurs données comparables.

Les sous-corpus sont représentatifs de divers domaines, types de discours et niveaux de langue et la répartition est la suivante : 10% de textes techniques, 14% de textes scientifiques, 11% de textes sociopolitiques, 13% de textes administratifs, 16% articles de journaux, 13% de textes littéraires, 13% de textes environnementaux, et 10% d'oral provenant d'enquêtes sociolinguistiques.

Afin de compléter les possibilités de recherche déjà offertes par la mise en réseau de ce corpus, il a semblé particulièrement intéressant de le soumettre à un traitement lexicométrique, permettant notamment de cibler certains caractéristiques sémantiques.

3. Méthodes et techniques

Le logiciel *Hyperbase* est un logiciel d'exploitation documentaire et de traitement quantitatif des grands corpus textuels qui autorise un ensemble de traitements sur des corpus de textes prédéfinis et numérisé. Il permet ainsi le traitement statistique et automatique des données, qui peut servir de plateforme pour diverses études sur un corpus défini ainsi que pour la comparaison avec les données du corpus *Frantext*.

Hyperbase permet non seulement l'exploitation documentaire et l'obtention des contextes et des concordances, mais de plus la distribution d'un mot peut être étudiée dans l'ensemble des textes qui composent le corpus de travail et visualisée grâce aux applications graphiques. L'exploration statistique du logiciel donne la possibilité d'analyses diverses, non seulement traditionnelles comme celles sur la richesse lexicale, l'accroissement lexical, la distance lexicale, la corrélation chronologique etc., mais plusieurs fonctions thématiques sont également disponibles. Le logiciel *Hyperbase* permet d'analyser les spécificités lexicales aussi bien d'un point de vue aussi bien endogène qu'exogène, il permet d'extraire les corrélats thématiques et recense aussi tous les termes situés dans l'environnement immédiat d'un mot donné, d'extraire le réseau isotopique. *Hyperbase* est aujourd'hui en mesure de traiter les sorties du lemmatiseur-étiqueteur *Cordial* pour la langue française et de *TreeTagger* pour l'allemand, l'anglais et l'italien. Il ne permet donc plus seulement traiter les "mots", mais aussi les lemmes, les codes grammaticaux, les enchaînements syntaxiques ou les corrélats sémantiques.

3. Analyses lexico-sémantiques

La thématique des différents récits peut être exploitée, avec le recours au logiciel logométrique *Hyperbase*, de façon très précise, en s'intéressant aux spécificités des différents récits, à leur évolution et à leurs contextes.

L'analyse des spécificités est une démarche classique, que le logiciel accomplit en s'appuyant sur *Frantext*, et plus précisément sur le corpus du XXe siècle. Les mots qui se trouvent en tête de liste mettent en relief la spécificité et la réalité quotidienne de la vie canadienne par rapport aux mots "franco-français" du corpus *Frantext*. Nous trouvons également des mots relatifs à la vie parlementaire et administrative, moins employés dans le corpus français, qui puise une large partie de ses textes dans la littérature.

Le logiciel permet également l'observation du vocabulaire spécifique de chacun des sous-corpus, c'est-à-dire une comparaison endogène. Cette spécificité est déterminée par le calcul de l'écart réduit pour chaque forme dans chaque partie du corpus. Les textes sont comparés, les uns après les autres, avec le corpus dans son ensemble. Ces comparaisons internes se justifient facilement, puisque le corpus est homogène et expressément conçu pour mettre en valeur les différences qui opposent les textes dans ce même ensemble. Les résultats sont très nets, ces mots reflètent parfaitement le genre du discours et nous donnent le profil caractéristique de chaque typologie. Les spécificités lexicales du genre scientifique nous affiche des lexèmes tels que : *modèle*, *carbone*, *réaction*, *électrons*, *température*, *silicium*, *équation* et *méthode*.

Le corpus administratif est caractérisé par des mots comme *article, employé, règlement, employeur, syndicat, leader, indemnité, motion* et *alinéa*, ce qui n'est pas très étonnant.

Dans le corpus "Environnement", il est souvent question de l'Europe avec les pays européens en tête de liste, tels *France, Paris, Italie, Espagne, Allemagne, Angleterre,* qui apparaissent avant des termes comme *vins*, *peinture* et *festival*.

Le corpus journalistique est caractérisé par des mots qu'appartiennent à ce monde : *illustration, photo, article* etc. , ainsi que des mots qui reflètent la vie courante canadienne comme *dollars*, *Montréal* et *Soleil.*

Rien d'étonnant à trouver des mots tels que *carbone, réactions, électrons, température* et *silicium* en tête de liste des spécificités du sous-corpus scientifique. En revanche, le parallélisme du corpus littéraire et du corpus oral l'est peut-être davantage.

Le discours des deux sous-corpus fait appel aux mêmes pronoms des deux premières personnes, aux démonstratifs *ça* et *ce*, aux verbes d'état ou courants comme *être, faire, dire* et *penser*. Il semble en effet que le discours littéraire contemporain présent dans notre corpus soit proche de l'oral, bien que les spécificités du langage familier québécois, comme *pis, ben, pron* et *faque*, soient moins représentées.

Le logiciel permet aussi l'observation de tous les mots présents dans l'entourage immédiat d'un mot choisi pour pôle, et la comparaison de la fréquence de ces corrélats dans ce sous-corpus, constitué par l'entourage immédiat, avec celle du corpus entier, pour ainsi extraire l'univers lexical ou bien le réseau isotopique qui entoure un mot.

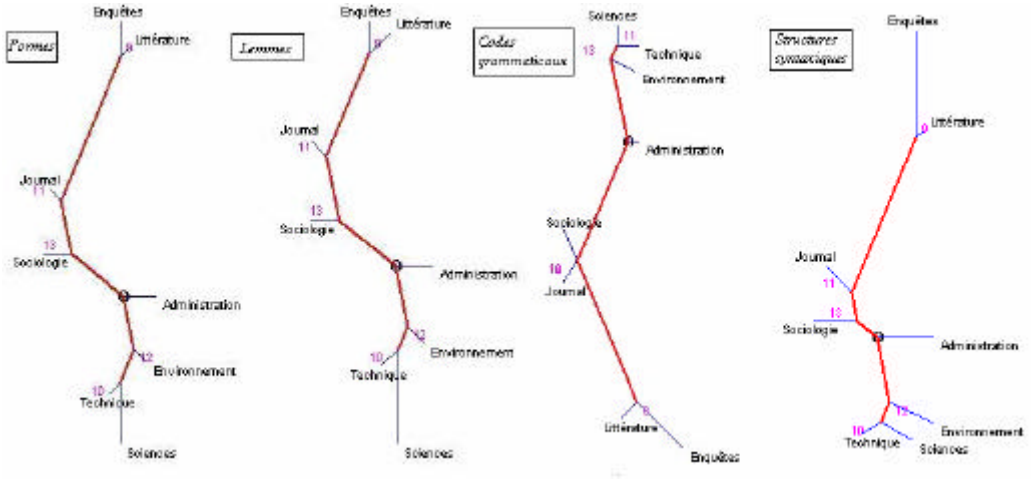
Il s'agit d'un calcul de spécificité particulier, puisqu'on ne recherche plus une relation entre un mot et un texte, mais une relation privilégiée entre les mots eux-mêmes - ce que mesure aussi le calcul de corrélation, quand deux séries sont juxtaposées.

Le clivage typologique est toujours présent et les résultats montrent des caractéristiques bien particuliers, spécifiques des différents genres.

De la même façon, l'analyse des corrélats sémantiques et thématiques révèle aussi les caractéristiques de chaque typologie présente dans ce corpus et l'analyse factorielle montre que les mêmes orientations des textes se retrouvent aussi bien aux niveaux lexical et syntaxique qu'au niveau thématique.

Une autre façon d'aborder genres et thèmes est l'étude de la distance lexicale ; il s'agit de considérer le vocabulaire intégral de chacun des textes du corpus et de rassembler ceux qui partagent des thèmes semblables. En réalité il existe plusieurs façons de faire ce calcul : on peut tenir compte ou non de la fréquence ; si on travaille sur V, on utilise la formule Jaccard, et on ne s'intéresse qu'à la présence ou à l'absence prenant appui sur les lemmes ou sur les graphies.

Les analyses arborées ci-dessous montrent les résultats du calcul non seulement sur le lexique, mais aussi sur les codes grammaticaux et sur les structures syntaxiques. On voit que les résultats sont sensiblement les mêmes, mettant en avant les mêmes oppositions génériques.



Graphique 1 : Analyse arborée de la distance intertextuelle, présence/absence V

On observe ici que l'on obtient la même image, alors que les objets considérés sont étrangers les uns aux autres : formes, lemmes, codes grammaticaux ou enchaînements syntaxiques.

Il est vrai qu'entre les formes et les graphies il existe une part commun, tout comme pour les codes et les structures syntaxiques, ces dernières étant des combinaisons des codes.

Mais quel lien existe-il entre les graphies et les structures ? On peut constater que les sous-corpus s'orientent différemment à tout niveau, lexical, grammatical ou syntaxique, selon le genre.

On effet, les variables que l'on croit indépendantes sont liées par des accords ou par une commune soumission à une très forte influence, celle du genre ou de la typologie, prépondérante dans toute analyse quantitative.

Chaque typologie a en fait son anatomie, sa physiologie et son fonctionnement, et cela transparaît très clairement dans ce corpus québécois.

Références

Adam J.-M. (2005). *Les textes types et prototypes : Récit, description, argumentation, explication et dialogue*, Paris, Arman Colin, collection Fac. Linguistique.

Brunet E. (2000). "Peut-on mesurer la distance entre deux textes ?", in Rastier F. (éd.) *Corpus littéraires – Recueil et numérisation, analyses assistées, didactique*,) Paris 20-21 octobre 2000.

Kastberg Sjöblom M. et Brunet E. (2000). "La thématique. Essai de repérage automatique dans l'œuvre d'un écrivain", in Rajman M. & Chappelier J.-C. (éds.), *JADT 2000, 5èmes Journées internationales d'Analyse statistique des Données Textuelles*, Ecole polytechnique fédérale de Lausanne, p. 457-465.

Kastberg Sjöblom M. 2002. *L'écriture de J.M.G. Le Clézio – Des mots aux thèmes*, Paris, Honoré Champion, 2006.

Malrieu D. & Rastier F. (2002). "Genres et variations morphosyntaxiques", in Angel Martin Municio (éd.), *Actas del segundo seminario de la escuela interlatina de altos estudios en lingüística aplicada, Matemáticas y tratamiento de corpus*, San Millán de la Cogolla, 19-23 septiembre de 2000, Logrono, Fundación San Millán de la Cogolla, p. 61-84.