Article

« Les bases de données textuelles et linguistiques à Sherbrooke : une banque en développement »

Pierre Martel et Michel Théoret

Revue québécoise de linguistique, vol. 20, n° 2, 1991, p. 123-142.

LES BASES DE DONNÉES TEXTUELLES ET LINGUISTIQUES À SHERBROOKE: UNE BANQUE EN DÉVELOPPEMENT

Pierre Martel et Michel Théoret

1. Introduction

Depuis quelques années, un courant de pensée largement répandu voudrait voir décrits, de façon systématique, la langue, et plus particulièrement, le vocabulaire des Québécois. Jusque-là, on avait plutôt tenté d'identifier les écarts entre la langue d'ici et le français standard; mais ce n'était qu'une étape.

Le lexique du français québécois doit être étudié non seulement dans les différences qu'il manifeste avec le français hexagonal, mais également dans sa totalité. L'équipe de Sherbrooke[1] a commencé ce travail dès 1971 en procédant, par des enquêtes sociolinguistiques, à l'étude globale du vocabulaire québécois. Cette recherche a permis, au moyen de corpus entièrement dépouillés et lemmatisés, d'établir, d'une part, une base de données textuelles et, d'autre part, une base de données linguistiques.

L'objectif principal de la présente recherche consiste à élaborer une base lexicographique québécoise. Pour décrire l'ensemble de la langue québécoise, nous sommes d'avis qu'il faut constituer une banque de textes oraux et écrits du français québécois. Il s'agit en fait de stocker des textes caractérisant la langue parlée et écrite de type littéraire, journalistique, administratif, technique, scientifique ou autre, en prenant soin de noter les particularités régionales, les données qui peuvent déjà apparaître comme archaïques, etc. Les données linguistiques seront indexées et lemmatisées dans le but de décrire les unités lexicales et les sens qui y seront attestés.

Nous exposerons dans les pages qui suivent la partie du travail déjà réalisé jusqu'ici et présentement en cours à Sherbrooke.

---

1. Il s'agissait, au point de départ, de Normand Beauchemin, directeur du projet de recherche, et de Pierre Martel. Quelques années plus tard, Michel Théoret s'est joint à l'équipe. Cette dernière comprend de plus maintenant Jean-Marcel Léard et Hélène Cajolet-Laganière.

2. Les corpus

2.1 Le corpus sociolinguistique de l'Estrie

Au cours des années 1971 à 1974, l'équipe de Sherbrooke a effectué une vaste enquête auprès de quelque cent cinquante informateurs et informatrices de l'Estrie (Beauchemin; 1972, 1977)[2]. Fondée sur les principes de la sociolinguistique, cette enquête a permis de recueillir un vaste échantillon de langue parlée, servant notamment à l'étude du lexique québécois. Cent sociotextes ont d'abord été retenus, transcrits et publiés (Beauchemin, Martel, Théoret, 1973 à 1981), puis entièrement lemmatisés (grâce à un lemmatiseur semi-automatique mis au point à Sherbrooke à la fin des années soixante-dix et sur lequel nous reviendrons plus loin). Sur le plan lexical, ces textes représentent un échantillon substantiel de la langue parlée en Estrie (plus considérable que le «français fondamental») (Gougenheim, 1964) et comprennent les éléments suivants:

- 363 234 mots-occurrences (N)
- 12 124 formes différentes
- 5 524 vocables différents (V) (sans les noms propres)

Les principaux résultats de cette étude ont déjà été diffusés (Beauchemin, Martel, Théoret, 1983), dans une publication où se retrouvent tous les mots (et leurs formes) relevés dans les cent textes en question, avec les indices statistiques nécessaires à leur analyse. Mais, puisque le traitement de ce corpus a été entièrement informatisé, il est possible aussi d'obtenir très rapidement toutes les informations utiles à des études plus poussées: des listes par fréquence croissante ou décroissante, par ordre alphabétique sur l'ensemble, une partie ou un seul des textes, par catégorie grammaticale; des listes d'anglicismes; des regroupements selon des critères sociolinguistiques; des analyses de corrélation (Martel, 1983), etc.

Ces textes ont été indexés, ce qui permet un accès direct à tout mot-occurrence avec son contexte (plus ou moins long selon la demande); on peut étudier ainsi la structure d'emploi de chacun et faire toutes les fines analyses sémantiques ou syntaxiques désirées.

2. Ce corpus a été recensé par Boisvert et Laurendeau (1988).

2.2 Le corpus élargi à l'ensemble du français parlé du Québec

Au cours des années quatre-vingt, nous avons étendu notre corpus initial de l'Estrie à celui de l'ensemble du Québec. Bien que, selon notre point de vue, le parler de l'Estrie renferme les principales caractéristiques linguistiques des diverses régions du Québec (Dulong et Bergeron, 1980), notre description, qui visait l'ensemble du français parlé du Québec, se devait de contenir des corpus provenant directement des principales régions du Québec et, notamment, des villes de Montréal et de Québec (qui contiennent environ les trois cinquièmes de la population québécoise). En ce qui concerne la «province», nous avons ajouté une autre région du Québec pour contrebalancer le poids de l'Estrie. La disponibilité du corpus du Saguenay-Lac-Saint-Jean nous offrait des données comparables aux autres, qui, de plus, étaient directement et facilement traitables.

Dans le but de compléter le portrait ainsi constitué du français parlé au Québec, nous aurions pu augmenter encore la taille des échantillons des corpus sociolinguistiques précédents. Cependant, comme certains d'entre eux n'abordaient qu'un nombre limité de thèmes, on a constaté que le vocabulaire, à partir d'une certaine taille, avait tendance à se répéter (en statistique, on appelle ce point le «point d'infléchissement» de la courbe asymptotique). Nous avons alors préféré tirer des échantillons d'un autre «genre» de textes en puisant dans les textes publiés qui reflètent la langue parlée.

Le deuxième sous-ensemble d'échantillons de langue orale a ainsi été tiré de textes publiés, mais dont on peut dire qu'ils ont été écrits pour être «dits». Ce sont les pièces de théâtre (exemple: «À toi, pour toujours, ta Marie-Lou» de Michel Tremblay), les téléromans (exemple: «Race de monde» de Victor-Lévy Beaulieu) et les textes radiophoniques (exemple: «Les mères dans l'histoire de l'humanité» de Carl Dubuc).

Nous avons enfin ajouté à ces trois types de langue, deux autres catégories de textes: ceux provenant de contes folkloriques (exemple: «La fille aux mains coupées» dans les *Archives de folklore*) et de monologues (exemple: «Ça dit qu'essa à dire» de Jacqueline Barrette). Ces deux dernières catégories de textes sont d'abord de la langue orale, mais ils sont aussi composés en vue d'une récitation publique (et éventuellement d'une publication). Le deuxième sous-ensemble de textes constituait

donc une sorte d'échantillon à mi-chemin entre la langue parlée réelle (des corpus sociolinguistiques) et la véritable langue écrite: nous avons appelé cet ensemble, celui des textes de «langue parlée non spontanée» ou de «langue parlée publiée».

En ajoutant ainsi une deuxième partie d'un demi-million de mots, nous voulions constituer un échantillon égal à celui provenant des corpus sociolinguistiques. L'ensemble du corpus d'un million de mots intégrant des genres de textes différents et des «situations» linguistiques différentes assurera, croyons-nous, une représentativité meilleure, et donc plus valable, du français parlé à notre époque au Québec.

Pour que l'ensemble du vocabulaire ainsi lemmatisé puisse être facilement analysable et que ses diverses parties puissent être facilement comparables, nous avons divisé ce corpus en deux sous-ensembles égaux, chacun composé de cinq tranches égales de 100 000 mots. C'est ainsi que nous avons:

1er sous-ensemble		*2e sous-ensemble*	
Estrie I	100 000	Contes	100 000
Estrie II	100 000	Monologues	100 000
Montréal	100 000	Théâtre	100 000
Québec	100 000	Textes radiophoniques	100 000
Saguenay-Lac-St-Jean	100 000	Téléromans	100 000
N:	500 000 mots	N:	500 000 mots
V:	7045	V:	9242

N:	1 000 000
V:	11 327
Formes:	35 927

Ces deux sous-ensembles sont évidemment exploitables de la même façon que le corpus publié de l'Estrie (voir 2.1). Ils sont disponibles sur support informatique (micro-ordinateur compatible IBM); tous les mots qu'ils contiennent peuvent être analysés en contexte à partir de l'ensemble, de l'un des sous-ensembles ou de l'une des tranches.

Ces listes feront d'ailleurs bientôt l'objet de la publication (actuellement chez un éditeur) d'un *Dictionnaire de fréquence des mots du français parlé au Québec*[3], qui comprend la liste alphabétique de tous les vocables accompagnés des formes attestées. Des listes classent également les mots par ordre de fréquence, de dispersion et d'usage. Enfin, une longue introduction présente les corpus, la façon dont ils ont été constitués, toutes les références des textes retenus, les informations utiles à la consultation du dictionnaire de même que quelques analyses montrant ce qu'il est possible d'en tirer.

2.3 Le corpus élargi aux textes littéraires du Québec

Les textes littéraires représentent le groupe de textes privilégiés dans toute banque servant à la description lexicale d'une langue. C'est particulièrement le cas en France (Frantext). La langue littéraire du Québec doit aussi faire l'objet d'un dépouillement systématique. Depuis deux ans (grâce à l'obtention de subventions *ad hoc*), nous avons entrepris la saisie d'un certain nombre d'oeuvres littéraires québécoises. Le choix des textes s'est fait jusqu'à maintenant en fonction de leur disponibilité sur un support intéressant pour nous et de leur intérêt immédiat. Le dépouillement de toutes les oeuvres importantes de la littérature québécoise se poursuit actuellement.

L'enregistrement informatique de ces textes s'est fait de la façon suivante:

- soit par saisie traditionnelle;
- soit par lecture optique et numéralisation;
- soit enfin par lecture directe de bandes de photocomposition informatisées fournies par certaines maisons d'édition.

2.4 L'état actuel de notre base de données textuelles

Grâce à l'ensemble de tous ces travaux, notre base de données textuelles de langue québécoise s'enrichit régulièrement: au moment d'écrire ces lignes, le nouveau corpus littéraire comprend plus de 710 000 mots, et ce nombre grandit sans cesse. Notre corpus de textes compte environ 2 300 000 mots, ce qui occupe un

3. Voir en annexe une page type de ce dictionnaire.

espace de près de 15 Méga-octets. Le nombre de formes différentes atteint environ 52 K. Tous ces textes ne sont pas encore lemmatisés, mais ils sont tous indexés et donc directement accessibles comme ceux du corpus de langue parlée; ils sont consultables indépendamment de ces derniers ou mis avec eux dans une banque d'une dimension intéressante reflétant bien la langue québécoise actuelle.

L'intérêt de notre banque de données, constituée au départ de divers corpus de langue orale, mais augmentée à l'heure actuelle de textes de la langue écrite et littéraire, ne fait pas de doute. En effet, il n'existe pas en ce moment au Québec une telle «banque de mots» comparable à la nôtre tant en ce qui a trait à son contenu qu'à la méthode utilisée (informatisée du début à la fin des opérations).

3. La méthodologie utilisée

3.1 Les textes (voir le schéma I)

Tout fonds de données lexicographiques suppose au préalable un choix difficile des textes à retenir (étape 0). Les textes oraux, tirés des grands corpus sociolinguistiques existant au Québec, constituent une partie importante des données de Sherbrooke[4]. La place accordée au corpus de l'Estrie (200 000 occurrences sur 500 000) tient principalement à des considérations pratiques. Les textes écrits ont été sélectionnés en fonction de critères linguistiques et selon leur accessibilité immédiate. Les textes écrits et oraux sont tous à la disposition des chercheurs, qui peuvent les consulter selon les besoins de leur recherche.

3.1.1 La saisie des textes

Sur le plan de la saisie du corpus, nous avons commencé par mettre en mémoire, de manière traditionnelle (dactylographie), l'ensemble des textes retenus avec les inconvénients (fautes de frappe notamment) que cette méthode comportait. Nous avons dû, de plus, mettre au point un système de reconnaissance des accents (nécessaire à la levée de bien des ambiguïtés) que les micro-ordinateurs de l'époque ignoraient.

4. Il existe une banque un peu semblable à la nôtre en Belgique et qui s'appelle *VALIBEL* (Variétés linguistiques du français de Belgique). Cette ***banque*** est gérée par une équipe de recherche attachée à l'Université de Louvain-la-Neuve (Francard, 1989).

Ce dernier point étant maintenant réglé, la lecture en est évidemment améliorée. Mais c'est surtout grâce à la lecture optique que la saisie se fait de façon beaucoup plus sûre et infiniment plus rapide. Les essais que nous avons faits à Sherbrooke sont concluants à cet égard. Pour peu que le texte soit imprimé de façon claire, le lecteur optique nous fournit une lecture numéralisée immédiatement «compréhensible» par le micro-ordinateur. Il demeure cependant quelques difficultés, par exemple pour distinguer les traits d'union lexicaux (comme dans les mots composés) des traits d'union typographiques (comme ceux en bout de ligne). Il ne reste alors qu'à indexer les textes, ce qui se fait de manière complètement automatique. On voit immédiatement le grand avantage, en termes de rapidité et d'efficacité, d'une telle méthode.

Pour les textes plus récemment édités par certaines maisons, les problèmes sont encore plus restreints. La lecture de bandes de photocomposition informatisées — nous en avons fait l'expérience — est plus sûre et moins coûteuse que la lecture optique, puisqu'elle permet de passer, moyennant des ajustements fort peu nombreux, directement à l'indexation des textes.

Nous croyons que cette étape est cruciale, car de la qualité de l'enregistrement (saisie) du corpus textuel dépendra la qualité de toutes les exploitations et des analyses qui suivront.

3.1.2 Le marquage des textes

Une intervention humaine (souvent appelée marquage ou balisage) doit avoir lieu au moment de la saisie des textes afin d'assurer une représentation extensive de nombreuses caractéristiques textuelles, c'est-à-dire de paramètres linguistiques (distinguer, notamment, les textes des enquêteurs de ceux des informateurs) et extralinguistiques (notes de l'éditeur ou, par exemple, les indications scéniques dans les pièces de théâtre). Ce marquage du texte est essentiel et il s'agit aussi bien d'indiquer le nom de l'auteur, l'année de l'édition, la pagination du texte que les variations dans les caractères d'imprimerie (italique, gras, majuscule, et les guillemets pour les citations et les mots étrangers). L'utilisation de ces données textuelles, leur interprétation par des consultants non spécialistes et l'objectif de certaines recherches nécessitent en effet que les mots soient situés dans leur contexte précis.

3.2 La base de données textuelles (BDT)

3.2.1 L'indexation et les concordances

Les textes informatisés doivent d'abord être indexés pour pouvoir être lus directement à l'écran ou imprimés. L'indexation est une opération consistant à doter chaque mot des coordonnées référentielles, appelées «adresses», qui permettent de le situer par rapport aux autres mots du texte et, subséquemment, de le retrouver. On peut demander alors n'importe quelle forme qui fait partie des textes informatisés et obtenir ainsi toutes les occurrences avec référence (texte, page, ligne) d'une forme quelconque. Celle-ci peut être lue dans un contexte de trois lignes ou plus et, si on le désire, de toute une page. Il est ainsi possible de produire de multiples concordances de tous les mots des textes enregistrés.

3.2.2 La lemmatisation

C'est évidemment grâce à l'emploi de la micro-informatique qu'a pu être traitée cette masse de documentation textuelle.

Dès 1977 ont été élaborés et mis au point les divers logiciels et le lemmatiseur semi-automatique nécessaire à l'analyse du corpus. Le fonctionnement de toute cette «mécanique» a déjà été expliqué de façon détaillée (Beauchemin, Théoret, 1984). Nous ne reviendrons ici que sur ses grandes caractéristiques. Une fois saisi, le texte est «lu» par la machine qui l'analyse mot à mot. Le lemmatiseur contient un dictionnaire, auquel 3500 formes de départ avaient été fournies; ce dictionnaire est toutefois dynamique, c'est-à-dire qu'il «apprend» les nouvelles formes rencontrées à mesure qu'on les lui donne, de sorte qu'il en reconnaît aujourd'hui plus de 35 000. Il y a trois niveaux à cette reconnaissance:

- le mot est connu et non ambigu: il est alors analysé, rattaché à son lemme et classé tout à fait automatiquement (ex. *plafonds* automatiquement rattaché au lemme *plafond* (subst. masc.).

- le mot est connu mais ambigu (*suis* de *être* ou *suis* de *suivre*, par exemple): le lemmatiseur affiche à l'écran les diverses possibilités et il suffit au linguiste d'appuyer sur la touche correspondant à l'un ou l'autre des emplois pour que l'analyse se fasse.

- l'analyse morpho-syntaxique est nécessaire au processus de désambiguïsation des unités de discours. Précisons que dans un grand nombre de cas, l'analyse de mots en principe ambigus se fait automatiquement grâce à une série de «règles de grammaire» contenues dans le lemmatiseur: ainsi la forme *est* sera-t-elle classée sous le verbe *être* si elle est précédée d'un pronom, sous le substantif *est* si elle suit un déterminant.

- le mot est inconnu: le linguiste doit fournir à la machine les éléments nécessaires: vocable, forme particulière, classe grammaticale, ambiguïté possible. La forme est alors classée et localisée (page, ligne) mais surtout retenue dans le dictionnaire dynamique qui l'analysera, la fois suivante, selon l'un des deux niveaux précédents.

3.2.3 La liste des vocables et les autres *index*

À la fin de ce travail étonnamment rapide, l'ordinateur nous fournit la liste alphabétique de tous les vocables, avec leur classe grammaticale, toutes les formes sous lesquelles ils ont été rencontrés et la fréquence de chacune, soit pour un texte donné, soit pour une tranche, un sous-ensemble ou le corpus total.

Ainsi, par exemple, on peut obtenir la liste de tous les verbes du premier groupe dans 1, 2, 3,... textes ou bien tel et tel verbe suivi de la préposition x ou y. On peut se procurer également le vocabulaire caractéristique d'un thème particulier (par exemple, le vocabulaire des «sucres»). Tous les produits fournis correspondent à des demandes portant sur des mots, des lexies, des collocations, des expressions syntagmatiques (*être + après + verbe*), des synonymes, des suffixes, etc. Les possibilités d'exploitation sont infinies et les demandes de consultation devraient s'accroître proportionnellement à la croissance de la BDT.

La convivialité du système rend son exploitation facile d'utilisation par les non-spécialistes. Dans notre cas, les postes de travail sont accessibles aux étudiants qui apprennent le mode d'emploi rapidement et facilement.

3.3 La base de données linguistiques (BDL) (voir le schéma II)

Cette deuxième étape est nécessaire dans la perspective d'une utilisation de la BDT à des fins dictionnairiques (Quemada, 1990). On sait par ailleurs que la

lemmatisation, qui permet de confectionner la BDL, devient vite une opération «coûteuse» (en ressources de toutes sortes) lorsqu'on veut l'appliquer à de très grands corpus (*Frantext*, par exemple).

Il n'est pas question d'examiner ici les hypothèses existantes qui permettent de procéder économiquement à une lemmatisation complète ou partielle des textes, mais nous voudrions insister uniquement sur son importance. En séparant les formes homographes de vocables différents et en regroupant les occurrences de toutes les formes fléchies d'une même unité lexicale, on passe en fait du niveau du discours à celui de la langue.

Les questions soulevées par l'opération de la lemmatisation sont fondamentales; ce sont celles principalement de l'opposition, non résolue théoriquement, entre polysémie et homonymie et celle de la délimitation d'un mot (qui varie entre l'unité minimale constituée par les caractères entre deux blancs et le syntagme lexicalisé du type *pomme de terre*, *chemin de fer*, etc.).

Toute recherche linguistique ne peut s'effectuer que sur des matériaux regroupés et non sur des formes éparpillées dans les textes. Comment en effet peut-on faire un travail d'analyse sur le mot *oeil* si on ne considère pas en même temps les occurrences du mot *yeux*? Comment présenter et analyser les sens et les emplois des verbes *placoter*, *poigner/pogner*, *maganer*, *faire*, etc., si toutes leurs formes morphologiques et leurs variantes orthographiques ne sont pas regroupées? Or, si certaines formes d'un même vocable se suivent dans l'ordre alphabétique des mots, il en est souvent autrement, ce qui présente une difficulté réelle. Dans notre BDT, on a ainsi trouvé les variantes *bean*, *bine* et *binne*, *j'ville*, *ch'fille* et *cheville*, *shop* et *chop* (pour le même vocable), etc.

L'identification des anglicismes, par exemple, qu'il faut considérer comme très importante dans l'analyse et le traitement de tout texte québécois (ne serait-ce qu'à une fin descriptive), est une raison supplémentaire justifiant l'identification des unités lexicales ou vocables des textes. Il est important à nos yeux de distinguer *tire* «bonbon», (*tire* «voiture» n'a pas été rencontrée), de *tire* «pneu», anglicisme. Dans notre BDL, nous avons identifié ce dernier par la marque (a).

3.4 La base de données dictionnairiques (BDD)

Avec la troisième étape, nous passons à l'analyse des matériaux dictionnairiques. Les données de la BDT (voir schéma II) et celles de la BDL fournissent alors au chercheur ou au rédacteur la matière nécessaire à sa recherche ou à la rédaction de textes langagiers ou dictionnairiques.

La liste exhaustive des vocables de la BDL, à laquelle s'ajoutent les index de fréquences (allant des mots les plus fréquents jusqu'aux hapax), permet d'arrêter, entre autres, la nomenclature du dictionnaire que l'on veut rédiger; cette nomenclature sera différente selon qu'il s'agit d'un dictionnaire fondamental, d'un dictionnaire pour les élèves du primaire ou, au contraire, d'un dictionnaire complet composé de plusieurs volumes.

La BDT permet particulièrement de sélectionner les exemples les plus significatifs et de citer les emplois les plus clairs et les plus courants. La conjonction des matériaux de la BDT et de la BDL permet évidemment le repérage des synonymes, des composés, des principales constructions, etc.

3.5 La rédaction des ouvrages

Il va sans dire que la rédaction des articles et des autres ouvrages a tout avantage à être effectuée sur micro-ordinateur. Un bon traitement de textes permet la composition dynamique des articles et rend les corrections faciles. On peut déjà envisager que les nouveaux textes dictionnairiques qui seront rédigés ainsi deviendront eux-mêmes des échantillons de discours (à titre de documentation métalinguistique et secondaire) qui seront aussi intégrés à la BDT.

Le deuxième niveau a été peu développé jusqu'ici à Sherbrooke, étant donné la taille encore limitée des deux bases. C'est quand même à partir de ces matériaux que notre collègue et coéquipier Normand Beauchemin a puisé l'information pertinente au choix de la nomenclature et à l'élaboration des 2300 articles qu'il a soumis aux responsables du *Dictionnaire du français plus.*

Nos deux bases de données contiennent cependant la partie fondamentale (vocabulaire de base) de la langue québécoise actuelle. Et ces bases de données continuent de grandir...

4. Quelques chiffres tirés du *Dictionnaire de fréquence des mots du français parlé au Québec*

4.1 Les mots les plus fréquents et les hapax

Malgré la somme intéressante des données déjà accumulées à Sherbrooke, il va de soi que son exploitation ne fait que commencer. À partir de notre BDL, nous avons esquissé le portrait statistique du vocabulaire du français parlé au Québec.

On constate que les 50 vocables les plus fréquents du corpus totalisent 63,9 % du texte total, soit 639 551 occurrences sur un million. Cette proportion s'élève à 66,1 % en moyenne pour les tranches de langue parlée et baisse à 61,79 % dans celles de la langue parlée non spontanée. Ainsi, les deux tiers d'un texte ou presque sont composés de quelques dizaines de mots dont moins d'une dizaine de verbes: *faire*, *aller*, *dire*, *savoir*, *voir* et *vouloir* en plus des deux auxiliaires *être* et *avoir*. Cela montre tout le soin qu'il faudrait accorder à l'apprentissage et à la parfaite maîtrise morphologique et sémantique de ces mots.

À l'autre extrême de l'échelle de fréquence, nous avons observé que l'effectif des vocables de fréquence 1 (les hapax) constitue 32 % de l'ensemble de tout le vocabulaire utilisé, soit 3646 unités lexicales sur 11 327. Cette proportion reste la même dans les échantillons de langue parlée, mais augmente à 36% dans ceux de langue parlée non spontanée. L'étendue du vocabulaire est également plus restreinte dans

4.2 Les anglicismes

Nous avons relevé tout vocable ou toute forme qui peut être considéré comme un anglicisme. Pour éviter le plus possible les discussions, nous avons suivi la norme du *Petit Robert* (1972, 1981) qui considère comme anglicisme (ou américanisme) tout «mot anglais (ou américain) employé en français et critiqué comme emprunt abusif ou inutile» (p.XXIV). C'est pourquoi (sans nous prononcer sur le bien-fondé des décisions des auteurs et bien que certains jugements soient, à cet égard, pour le moins discutables), nous avons considéré comme anglicismes tous les mots inscrits dans le *Petit Robert* et portant cette marque d'usage: ainsi *mass-média* n'est pas un anglicisme non plus que *knock-out*, mais *flash* et *juke-box* le sont.

Pour les anglicismes utilisés seulement au Québec, la tâche n'était pas facile. La plupart ne posent pas de problème (*fun*, *kick*, etc.). Mais certains font difficulté: *malle* au sens de «courrier» et *rainette* au sens de «couvre-chaussures en caoutchouc» sont-ils, par exemple, des anglicismes ou des régionalismes québécois à partir de mots français ou dialectaux du galloroman? Dans tous ces cas, nous avons suivi, ou bien les indications du *Glossaire* (Glossaire, 1968) (mais ses étymologies ne sont pas toujours sûres) ou bien notre sentiment et nos connaissances linguistiques.

Les anglicismes sont au nombre de 2861 occurrences sur le million de mots que comprend notre corpus de langue parlée. La fréquence moyenne des anglicismes en français parlé au Québec ne serait donc, si on en juge à partir de cet échantillon, que de deux sur mille mots[5]. En ce qui a trait au vocabulaire, nous avons dénombré 699 anglicismes sur les 11 327 vocables du corpus (ce qui représente 6% de l'ensemble).

À ces premières données, on peut ajouter la distribution de leur fréquence et y rattacher chacun des effectifs rencontrés:

5. Nous n'avons pas cependant noté tous les anglicismes de sens (ou calques), fréquents au Québec, mais dont l'analyse est très souvent sujette à controverse.

Fréquence	V anglicismes
1	326
2	131
3	64
4	45
5	30
6	18
7	16
8	11
9	9
10	7
11	5
12	3
13	1
14	1
15 (≥)	32
Total	699

Il y a donc lieu de constater que 93% d'entre eux apparaissent moins de dix fois (650 sur 699). Ceux qui ont une fréquence supérieure à 14 sont les suivants, par ordre de fréquence décroissante:

Anglicismes (fréquence)			
fun	(171)	fournaise	(26)
job	(125)	smart	(25)
chum	(119)	steak	(23)
party	(62)	truck	(23)
gas	(41)	jean	(22)
shop	(41)	toffer, verbe	(22)
pan	(40)	tough	(20)
sleigh	(36)	steady	(19)
foreman	(30)	break	(18)
jobber, subst.	(29)	clairer	(18)
runner, verbe	(29)	feeler	(18)
watcher,verbe	(29)	tague	(17)
shower, subst.	(28)	hell	(16)
peanut	(27)	rubber, subst.	(16)
show	(27)	slacker,verbe	(16)
can	(26)	track	(15)

Ces quelques données quantitatives permettent certainement d'avoir une vue un peu plus précise du phénomène de l'emprunt au Québec et aideront, croyons-nous, à mieux cibler les interventions de nature prescriptive.

5. Conclusion

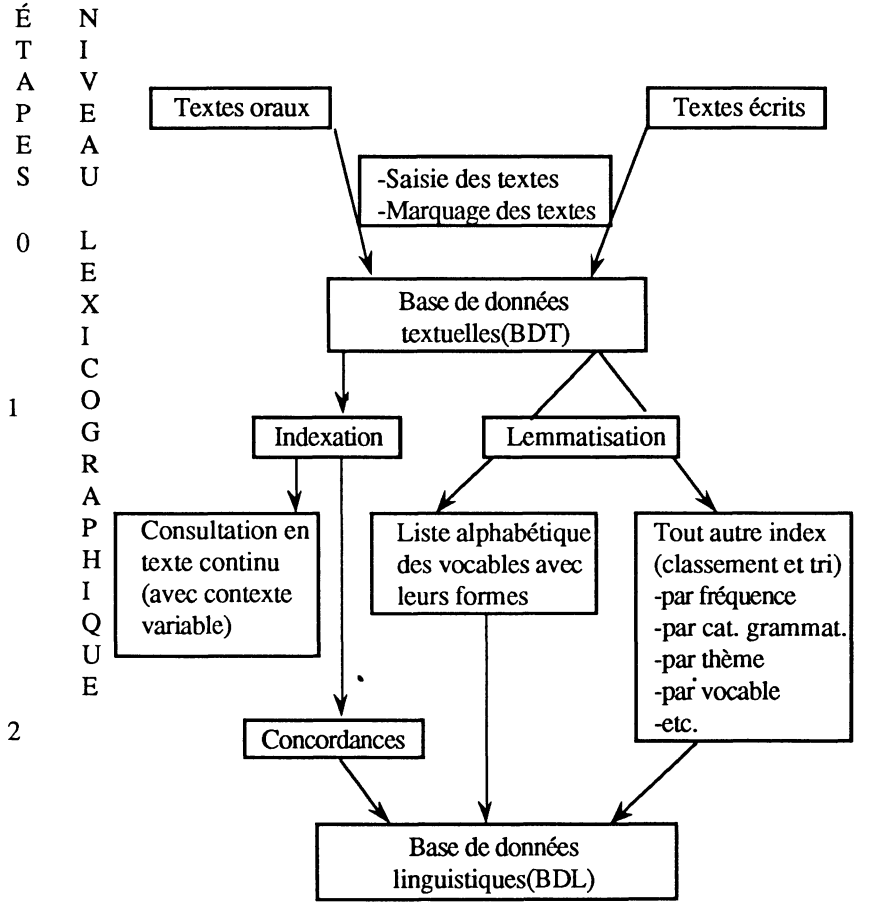
Bénéficiant maintenant de l'expérience que nous avons développée à Sherbrooke depuis plusieurs années, nous poursuivons notre travail en mettant l'accent principalement sur deux dimensions.

D'abord, nous voulons augmenter nos banques de données textuelles et linguistiques. L'échantillon actuel que nous possédons de la langue parlée réelle nous paraît suffisant pour refléter le français oral tel qu'il se dégage des corpus sociolinguistiques du Québec. Il nous faut par contre poursuivre la saisie de textes littéraires et surtout, comme le propose le Conseil de la langue française dans son dernier avis sur *L'aménagement de la langue: pour une description du français québécois* (Conseil, 1990), diversifier davantage les textes de notre banque. L'introduction de nouveaux textes, de nature et de thèmes fort éloignés des précédents, augmentera rapidement et sensiblement le nombre de formes et de vocables nouveaux du français québécois standard.

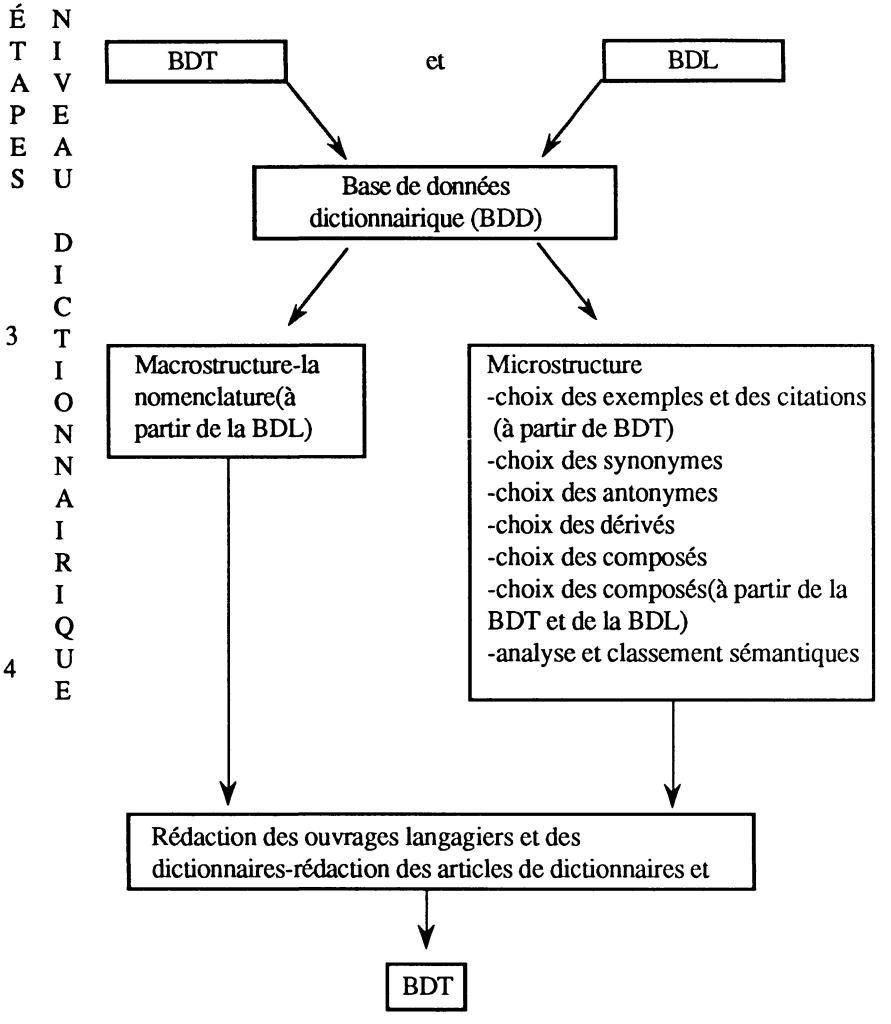
D'autre part, la taille de nos bases de données est suffisante pour permettre un certain nombre d'études de contenu. Notre équipe, qui s'agrandira vraisemblablement au cours de la prochaine année, a commencé des analyses sémantiques et morpho-syntaxiques du français québécois. Modeste au départ, la banque de données textuelles et linguistiques de Sherbrooke, grâce à l'augmentation de ses données et de ses ressources, apportera certainement une contribution importante au futur fonds québécois de données linguistiques. Elle permettra surtout de constituer, à partir des données de corpus réels, la base lexicographique québécoise tant souhaitée par les Québécois.

Pierre Martel et Michel Théoret
Université de Sherbrooke

LES BASES DE DONNÉES TEXTUELLES
ET LINGUISTIQUES
SCHÉMA I



LES BASES DE DONNÉES TEXTUELLES
ET LINGUISTIQUES
SCHÉMA II



Cl	Vocable/formes		Est.1	Est.2	Mont.	Que.	Slsj.	Tot1	Cont.	Thtr.	Mnlg.	Rad.	Tlrm.	Tot2	G.tot.	Disp.	Usage
2	**main(a)**		-	-	-	-	-	-	-	-	1	-	-	**1**	1		
2	**main-d'oeuvre**		-	2	-	1	1	4	-	1	-	-	-	**1**	5		
6	**maintenant**		4	53	25	26	3	111	9	33	25	35	52	**154**	265	**78.82**	208.8
		zp=	-4.32	7.31	0.66	0.90	-4.56		-4.39	0.44	-1.17	0.85	4.27				
		zt=	-4.61	5.43	-0.31	-0.10	-4.81		-3.58	1.33	-0.31	1.74	5.22				
	maintenant		4	53	25	26	3	111	9	33	25	35	41	**143**	254		
	maint'nant		-	-	-	-	-	-	-	-	-	-	11	**11**	11		
1	**maintenir**		1	1	-	-	-	2	-	-	-	1	1	**2**	4		
	maintenir		-	-	-	-	-	-	-	-	-	-	1	**1**	1		
	maintiennent		-	1	-	-	-	1	-	-	-	-	-	-	1		
	maintiens 1		-	-	-	-	-	-	-	-	-	1	-	**1**	1		
	maintient		1	-	-	-	-	1	-	-	-	-	-	-	1		
2	**maintien**		-	-	-	-	4	4	-	-	-	-	-	-	4		
2	**maire**		-	-	-	-	2	2	2	50	5	35	1	**93**	95	**40.70**	38.6
		zp=	-0.71	-0.71	-0.71	-0.71	2.83		-4.30	8.14	-3.53	4.25	-4.56				
		zt=	-3.25	-3.25	-3.25	-3.25	-2.56		-2.56	>+9.99	-1.54	8.72	-2.91				
	maire		-	-	-	-	2	2	2	50	5	20	1	**78**	80		
	maires		-	-	-	-	-	-	-	-	-	13	-	**13**	13		
	pro-maires		-	-	-	-	-	-	-	-	-	2	-	**2**	2		
2	**mairesse**		-	-	-	-	-	-	-	10	-	-	-	**10**	10		
2	**mairie**		-	-	-	-	1	1	-	-	-	-	-	-	1		
6	**mais**		726	949	763	992	527	3957	374	526	549	723	733	**2905**	6862	**91.03**	6246.4
		zp=	-2.60	6.26	-1.13	7.97	<-9.99		-9.60	-2.55	-1.48	6.59	7.05				
		zt=	1.60	>+9.99	3.09	>+9.99	-6.41		<-9.99	-6.45	-5.52	1.48	1.88				
	main		-	-	-	-	1	1	-	-	-	-	-	-	1		
	mais		726	949	763	992	525	3955	374	526	548	723	733	**2904**	6859		
	mè'		-	-	-	-	1	1	-	-	1	-	-	**1**	2		
2	**maison**		117	87	144	106	87	541	64	69	115	141	102	**491**	1032	**91.66**	945.9
		zp=	0.95	-2.28	3.85	-0.24	-2.28		-3.86	-3.29	1.90	4.83	0.43				
		zt=	1.43	-1.68	4.23	0.29	-1.68		-4.07	-3.55	1.22	3.92	-0.12				
	maison		103	77	111	92	80	463	63	69	103	127	101	**463**	926		
	maisons		14	10	33	14	7	78	1	-	12	14	1	**28**	106		
2	**maisonnée**		-	-	-	-	-	-	-	4	-	-	1	**5**	5		
2	**maitre**		-	3	3	3	2	11	24	3	1	5	4	**37**	48	**54.59**	26.2
		zp=	-1.66	0.60	0.60	0.60	-0.15		6.82	-1.81	-2.63	-0.99	-1.40				
		zt=	-2.31	-0.87	-0.87	-0.87	-1.35		9.24	-0.87	-1.83	0.10	-0.38				
	maitre		-	3	3	3	2	11	24	3	1	5	3	**36**	47		
	maitres		-	-	-	-	-	-	-	-	-	-	1	**1**	1		
2	**maitresse**		7	1	1	4	5	18	11	-	3	13	1	**28**	46	**69.22**	31.8
		zp=	2.00	-1.53	-1.53	0.24	0.82		2.55	-2.65	-1.23	3.50	-2.17				
		zt=	1.18	-1.77	-1.77	-0.29	0.20		3.15	-2.26	-0.79	4.13	-1.77				
	maitresse		7	1	1	3	1	13	11	-	3	13	1	**28**	41		
	maitresses		-	-	-	1	3	4	-	-	-	-	-	-	4		
	metresses		-	-	-	-	1	1	-	-	-	-	-	-	1		
2	**maitrise**		-	-	-	2	-	2	-	1	-	-	-	**1**	3		
2	**majesté**		-	-	-	-	-	-	-	-	-	8	-	**8**	8		
	majesté		-	-	-	-	-	-	-	-	-	7	-	**7**	7		
	majestés		-	-	-	-	-	-	-	-	-	1	-	**1**	1		
3	**majeur**		-	-	1	2	1	4	-	5	-	-	1	**6**	10	**50.56**	5.0
	majeur		-	-	1	-	-	1	-	1	-	-	-	**1**	2		
	majeure		-	-	-	2	1	3	-	4	-	-	1	**5**	8		
3	**major**		-	-	-	-	-	-	1	-	-	-	-	**1**	1		
	majors (subst)		-	-	-	-	-	-	1	-	-	-	-	**1**	1		
2	**majorette**		-	-	-	-	-	-	-	1	-	-	-	**1**	1		
	majorettes		-	-	-	-	-	-	-	1	-	-	-	**1**	1		
2	**majorité**		7	7	4	4	-	22	-	3	7	-	1	**11**	33	**71.41**	23.5
3	**majuscule**		-	-	-	-	-	-	-	-	-	2	-	**2**	2		

Références

BEAUCHEMIN, Normand (1972) *Le questionnaire*, Document de travail n°1, Université de Sherbrooke, 46 pages.

BEAUCHEMIN, Normand (1972) *Quelques traits de prononciation québécoise dans un contexte anglophone qui les influence?*, Document de travail n°2, Université de Sherbrooke, 30 pages.

BEAUCHEMIN, Normand (1972) *La diphtongaison en Estrie*, Document de travail n°4, Université de Sherbrooke, 32 pages.

BEAUCHEMIN, Normand (1977) *Données sociologiques*, Document de travail n°11, Université de Sherbrooke, 53 pages.

BEAUCHEMIN, N. et P. Martel (1973, 1975, 1977 et 1978) *Échantillons de textes libres*, n^{os} I, II, III et IV, Documents de travail n°8, n°9, n°10, n°12, Université de Sherbrooke.

BEAUCHEMIN, N., P. Martel et M. Théoret (1980-1981) *Échantillons de textes libres*, n^{os} V et VI, Documents de travail n°16 et n°17, Université de Sherbrooke.

BEAUCHEMIN, N., P. Martel et M. Théoret (1983) *Vocabulaire du québécois parlé en Estrie, Fréquence-dispersion-usage*, Document de travail n°20, Université de Sherbrooke, 303 pages.

BEAUCHEMIN, N. et M. Théoret (1984) «Micro-Solivo: un lemmatiseur semi-automatique pour le québécois parlé», dans *Revue de l'association québécoise de linguistique*, vol.3, n°3, pp.19-38.

BOISVERT, Lionel et Paul Laurendeau (1988) «Répertoire des corpus québécois de langue orale» dans *Revue québécoise de linguistique*, Université du Québec à Montréal, Montréal, volume 17, n°2, pp.241-262.

Conseil de la langue française (1990) *L'aménagement de la langue: pour une description du français québécois*, Québec, 35 pages.

Dictionnaire du français plus (1988) Centre éducatif et culturel inc., Montréal.

DULONG, Gaston et Gaston Bergeron (1980) *Atlas linguistique de l'Est du Canada*, Éditeur officiel du Québec, 10 tomes.

FRANCARD, Michel (1989) «VALIBEL: la première banque de données sur les variétés orales du français de Belgique», dans *Travaux de linguistique*, 18, pp.165-168.

Frantext, Banque de données textuelles de l'INaLF, C.N.R.S., Nancy, France.

Glossaire du parler français au Canada (1968) Québec, Presses de l'Université Laval, (réimpression de l'édition de 1930).

GOUGENHEIM, G., R. Michéa et P. Rivenc (1964) *L'élaboration du français fondamental (1er degré)*, Paris, Didier, 302 p.

MARTEL, Pierre (1983) «Les variables lexicales sont-elles sociolinguistiquement intéressantes?», dans les *Actes du XVIIe Congrès international de linguisique et philologie romanes*, Aix-en-Provence, volume 5, pp.183-193

Petit Robert, (1972 et 1981) Société du Nouveau Littré, Paris.

QUEMADA, Bernard (1990) «La nouvelle lexicographie», dans *La lingüistica Aplicada* Publicacions Universitat de Barcelona, Barcelona, pp.55-78.