Analyser les entretiens sociologiques

Jean-Guy Bergeron, Dominique Labbé

Analyser les entretiens sociologiques

Jean-Guy Bergeron [1], Dominique Labbé[2]

1. Ecole des relations industrielles - Université de Montréal - Montréal

(jean-guy.bergeron@umontreal.ca)

2. CERAT-IEP - Université Pierre Mendès-France - Grenoble

(dominique.labbe@iep-grenoble.fr)

Publié dans : Purnelle Gérald, Fairon Cédrick et Dister Anne (Eds). *Le poids des mots. Actes des 7e journées internationales d'analyse des données textuelles*. Louvain-la-Neuve : Presses Universitaires de Louvain, 2004, p. 136-147.

Abstract

How to analyse interviews? We present a collection of 61 interviews about industrial relations in some firms in Quebec. How to transliterate speeches? standardise spellings? mark up the text? tag every word… Then these normalised and tagged texts must be compared with every day speech of the whole population. An experiment is presented with the help of more than 300 interviews, held with French people, the total of which exceeds 2 millions words. It appears that the Quebeckers favour nominal groups and that they use less adverbs and conjunctions than French people whose speeches seem to be rather tense.

Résumé

On examine les problèmes posés par l'analyse des entretiens sociologiques à l'aide d'un corpus d'une soixantaine d'interviews à propos des relations industrielles dans les entreprises du Québec : normalisation orthographique, balisage et lemmatisation des textes. Il faudrait pouvoir comparer les enquêtés avec la population générale. Nous donnons un exemple de cette démarche à l'aide d'un corpus de plus de 300 entretiens réalisés avec des Français (plus de 2 millions de mots). Il apparaît que les Québécois préfèrent le groupe nominal et qu'ils utilisent moins d'adverbes et de conjonctions et que leurs propos sont nettement moins tendus que ceux des Français.

Mots-clefs : entretiens sociologiques - France - Québec - lemmatisation - corpus representatif - statistique lexicale.

L’entretien est l’outil privilégié des sociologues. Sans être totalement empirique, l’analyse des transcriptions est encore largement dominée par les méthodes qualitatives. On en trouvera un exemple dans l'ouvrage de Demazière et Dubar (1997) et dans le “symposium” organisé, autour de cet ouvrage, par la revue *Sociologie du Travail* en 1999. Mais l’idée d’une formalisation nécessaire et d’une mise à l’épreuve des outils tend à s’imposer (Jenny, 1997).

Depuis une douzaine d'années, le premier signataire de cette communication utilise, pour l'analyse de ses entretiens, des logiciels d'analyse du contenu comme “Nu-dist”. Par exemple, il a conduit un dépouillement de ce genre sur une soixantaine d'entretiens réalisés, de 1995 à 1998, dans le cadre d'une étude sur la négociation collective raisonnée au Québec (Bergeron & Bourque, 1996, 1998, 2003). Mais il est évident que, si les méthodes “caqdas” (Computer Assisted Qualitative Data Analysis Software) introduisent de la rigueur, l'opérateur continue à utiliser une grille d'analyse largement *a priori* — avec le risque d'introduire des biais — et qu'il peut aussi passer à côté de choses importantes qu'il n'aura pas pensé à coder… Les méthodes quantitatives de la statistique lexicale, peuvent-elles suppléer certaines de ces carences et permettre aux sociologues de porter un regard plus objectif sur leurs textes ? C'est ce que nous allons montrer à l'aide de ces 61 entretiens — impliquant au total plus de 70 locuteurs sans compter les enquêteurs. On examinera d'abord les “coûts” — c'est-à-dire, essentiellement, les contraintes qui pèsent sur la saisie des enregistrements — avant de poser le problème des étalons de référence et des méthodes de calcul. Enfin, on évoquera succinctement la richesse des résultats.

I. Saisie, balisage, étiquetage des corpus. Etalons de comparaison

Norme de saisie

Il n’existe pas à l’heure actuelle de véritable “code” de la transcription de l’oral. On a proposé des règles spécifiques visant notamment à restituer le rythme du propos (Benveniste & JeanJean, 1987). Bien que développée dans un laboratoire important et publiée il y a près de vingt ans, cette norme ne s’est pas imposée, car elle est complexe et éloignée des conventions usuelles. Comment la faire accepter par les opérateurs qui sont chargés de la saisie de ces entretiens mais à qui l'on demandera de continuer à utiliser des conventions profondément différentes pour le reste de leur travail ?

Ne vaut-il pas mieux utiliser la norme standard quitte à accepter une certaine perte d'information ? Ce standard ne peut être que la norme “sténotypique”, telle qu'elle est enseignée dans les écoles de secrétariat. Elle consiste, dans son principe, à faire entrer le mieux possible l'oral dans le lit de Procuste de l'écrit avec d'évidentes faiblesses comme pour la ponctuation où les résultats peuvent varier grandement selon l'opérateur. Si celui-ci privilégie la scansion au détriment de la syntaxe : la virgule représente une brève interruption ; une interruption plus longue est transcrite grâce au point si elle est précédée d'une baisse de l'intonation et par un point d'exclamation ou d'interrogation lorsque l'intonation ou la syntaxe le suggèrent. Si l'opérateur privilégie l'analyse syntaxique et sémantique, il ponctuera en fonction des périodes oratoires, en ayant recours aux trois points (souvent envahissants) quand la période ne se termine pas logiquement (du moins à ses yeux).

D'autres inconvénients paraissent inévitables. Signalons l'exemple frappant du participe passé avec l'auxiliaire “avoir”. Dans leurs propos spontanés, ou dans la vie quotidienne, la plus grande partie des locuteurs négligent l'accord (pour les verbes du 2e et du 3e groupes, le féminin doit s'entendre ; de même pour la liaison quand le 's' du pluriel est suivi par un mot commençant par une voyelle). Eh bien ! les secrétaires corrigent systématiquement et sans en avoir même conscience. Dans le cas d'une exploitation secondaire, on ne dispose généralement pas des bandes magnétiques, de telle sorte que sur ce point — crucial pour la réforme de l'orthographe —, ces corpus ne peuvent renseigner sur l'usage “réel” du français...

Enfin, sur ce premier point, une normalisation des graphies est indispensable, notamment pour les sigles, les abréviations, les noms propres, spécialement les patronymes et les toponymes étrangers. Cette tâche est partiellement effectuée par des automates, mais les interventions manuelles sont nécessairement nombreuses et doivent suivre des règles bien précises…

Balisage

Quand on consulte un corpus d’entretiens saisis par plusieurs opérateurs différents, on constate toujours que l'identification des questions et des réponses n’est pas stable. Par exemple, les questions sont parfois en italiques, parfois précédées de “E :” (“enquêteurs”), quelquefois même sans aucun identifiant… On introduit donc des balises qui délimitent les séquences du texte : en-têtes, remarques, questions et réponses (voir Labbé, 2001 et Labbé, 2002). Grâce à ces balises, l’opérateur pourra isoler le texte des réponses — c’est ce que nous ferons dans la suite de cet exposé — ou celui des questions, s’il s’intéresse au style de la sociologie… etc. Cette opération ne supprime rien, au contraire, elle facilite le traitement de l'information recueillie.

Lemmatisation

La plupart des mots sont susceptibles d'avoir plusieurs graphies : majuscules ou minuscules, élisions, abréviations… Dans certains entretiens, “monsieur” est écrit en toutes lettres, ailleurs M (suivi ou non d'un point), dans d'autres encore : Mr… Les noms de lieux et de personnes sont transcrits phonétiquement (si l'on n'a pas pris la précaution de les transmettre auparavant à l'opératrice). Les chiffres et les dates sont d'une infinie variété, parfois en lettres, parfois en chiffres, avec une virgule comme séparateur de millier, un blanc ou rien du tout. Par exemple, 1990, '90 mais aussi, plus insidieusement : “I99O” (sur le clavier, les I et O majuscules sont plus facilement accessibles que les chiffres…). Plus généralement, des milliers de mots français ont plusieurs graphies (événement et évènement ; puis et peux, etc.) et la réforme facultative de l'orthographe a encore considérablement augmenté leur nombre.

A l'inverse, des mots différents s'écrivent de la même manière. Par exemple, “je suis” (suivre ou être ?), “l'est” (article + nom ou pronom + verbe ?) ou “prise(s)” : substantif féminin (prise de courant), adjectif “pris” (au féminin), verbe “prendre” au participe passé ou… verbe “priser”. Dans tout texte français, ces homographies touchent plus du tiers des mots.

L'énoncé de ces problèmes contient les solutions : normaliser les graphies (un mot, une seule orthographe) et attacher à chaque mot une étiquette qui l'identifie complètement (entrée de dictionnaire et catégorie grammaticale). C'est en s'inspirant de cette idée que,

il y a une quinzaine d'années, a été mise au point une chaîne de traitement du français contemporain (Labbé, 1990). La nomenclature des mots, apprise à l'ordinateur, est systématique (par exemple, en français, les substantifs se distinguent par le genre, donc tous les substantifs doivent se voir affecter le masculin ou le féminin), elle est exhaustive (tous les mots doivent y trouver leur place), elle exclut tout double compte, elle ne comporte pas de catégorie ad hoc, ou fourre-tout, etc. Le principe général consiste à regrouper les flexions d'un même mot sous une forme vedette (“lemme”) auquel est associée une catégorie grammaticale. Ainsi, les conjugaisons d'un même verbe sont groupées sous son infinitif ou les pluriels du substantif sous le singulier ou encore les féminins et pluriels de l'adjectif sous le masculin singulier. Par exemple, “être v.” regroupe toutes les formes conjuguées de ce verbe, tandis que “être n. m.” ne se rencontre que sous le singulier et le pluriel.

Remarquons enfin que la lemmatisation doit être réversible — c'est-à-dire qu'on peut retrouver le texte original, sans altération, à partir du fichier des lemmes et qu'elle ne doit pas comporter d'erreur. Telle est la raison pour laquelle les automates — élaborés il y a 15 ans pour la normalisation et la lemmatisation par D. Labbé — résolvent en moyenne 99% des problèmes, laissant à l'opérateur les quelques cas douteux qu'une chaîne entièrement automatique ne pourrait traiter sans erreur…

Les bénéfices de ces opérations sont multiples. Par rapport aux traitements sur les formes graphiques brutes, la normalisation et la lemmatisation redonnent une existence aux verbes (en rassemblant leurs multiples flexions sous une étiquette commune). On peut retrouver certains mots comme le point cardinal “est”, les substantifs “être”, “avoir”, “avions”… dont les occurrences sont habituellement noyées dans l'océan des formes verbales homographes. Au-delà de ces avantages, la normalisation et la lemmatisation rendent possibles de nombreuses opérations statistiques dont cette étude donnera quelques exemples. En premier lieu, on peut comparer toutes sortes de corpus entre eux.

A quoi comparer ?

Si tous les entretiens ont été saisis, balisés et étiquetés en suivant rigoureusement la même norme, ils deviennent comparables entre eux. On peut opérer des classifications qui feront apparaître les principales sous-populations et mettront en lumière les caractéristiques, de vocabulaire ou de style, propres à chacun de ces groupes (pour un compte rendu de ces opérations sur le corpus “négociation raisonnée”, cf. Bergeron & Labbé, 2000). La limite est évidente : on voit ce qui différencie les sous-groupes mais il est plus difficile de révéler ce qui les unit. Comment retrouver les caractéristiques, communes à l'ensemble des individus interrogés, qui les singulariseraient par rapport au reste de la population ? Il faudrait pour cela disposer d’un étalon de référence, une vaste collection d'échantillons représentatifs des pratiques langagières dans la population générale. Un tel étalon existe pour beaucoup de langues. Historiquement, le *British National Corpus* est le premier apparu au début des années 1990 (voir le numéro 8-4 (1993) de *Literary & Linguistic Computing* et Burnard, 1995). Sur l'oral, voir les articles de : Crowdy, 1993 et Nelson, 1997. Les derniers corpus représentatifs parus concernent le tchèque (Kucera, 2002) et l'écossais (Douglas, 2003).

Il n'existe rien de tel pour le français. Certes, la définition d’un étalon de référence pose de multiples problèmes (voir Biber, 1993). Par exemple, a priori, il y a autant de

manière de parler français que de territoires francophones. Non seulement, il faudrait avoir des corpus représentatifs du français parlé, en Belgique, au Canada, en France, au Sénégal, en Suisse… mais ces corpus nationaux devraient aussi comporter des sous-ensembles pour ne pas négliger les différences de parler entre Bruxelles, la Wallonie ou entre le Québec, l'Ontario, le Nouveau Brunswick… Pour l'instant, le seul embryon existant a été réalisé par l'Université de Sherbrooke (Centre d'analyse et de traitement informatique du français québécois), et ce corpus n'est que partiellement étiqueté (www.userb.ca/Catifq/bdts).

Comme on le pressent, ces outils seraient très utiles (voir par exemple : Habert et Al, 1997). Nous allons le suggérer en utilisant pour cela un corpus "de référence" qui n'a pas de prétention à la représentativité. Il s'agit d'une vaste base de données rassemblant toutes les transcriptions de l'oral qui nous ont été confiées, depuis une dizaine d'années, aux fins de lemmatisation et de traitement (voir en annexe, une présentation de ce corpus nommé dans la suite de cette communication : "français oral"). Naturellement, on ne peut tirer de telles expériences que des inférences limitées. Il s'agit plutôt de se rendre compte des difficultés que rencontrerait une entreprise comme celle du BNC, sur le français (enregistrement, transcription, correction, étiquetage, indexation et traitement…). Par exemple, nous n'avons pas évoqué, faute de place, les problèmes juridiques posés par l'entreprise (protection de l'anonymat des personnes, de leur vie privée, de la propriété intellectuelle…). Il s'agit aussi d'avoir un aperçu de ce que pourraient apporter les corpus représentatifs pour les grammairiens, les lexicographes, les traducteurs, les enseignants, les gestionnaires de bases de données, etc.

Nous allons en donner deux exemples : la comparaison des catégories grammaticales et l'étude du vocabulaire caractéristique du corpus "négociation raisonnée".

II. Comparaison des catégories grammaticales

L'étiquetage systématique des textes rend enfin possible l'étude des parties du discours. En effet, la densité des verbes, des noms ou des mots outils varie en fonction des locuteurs et elle est sensible aux thèmes abordés. Le tableau ci-dessous résume les principaux résultats de la comparaison entre le corpus "négociation raisonnée" et le corpus "français oral". Dans le cas précis, la comparaison soulève une question supplémentaire : les différences proviennent-elles de spécificités propres au français du Québec ? Par exemple, l'excédent des "mots étrangers" peut facilement être rattaché à la situation particulière du pays et à l'emploi, dans la conversation courante, d'un nombre significatif de mots anglais. Ces emprunts ont fait l'objet d'une étude de l'Université de Sherbrooke (www.userb.ca/Catifq/angliweb). Notre corpus en apporte de nombreuses illustrations dont l'évocation dépasserait le cadre de cette communication.

Ecarts dans les densités d'emplois des principales catégories grammaticales entre le corpus “Négociation raisonnée” et le corpus de référence “Français oral”.

Catégories	Densité des catégories dans le sous corpus	Comparaison avec le français oral
Verbes	**19.8**	**+2.6**
Formes fléchies	*12.7*	*-7.4*
Participes passés	*3.4*	*+37.2*
Participes présents	*0.1*	*+35.8*
Infinitifs	*3.5*	*+19.1*
Noms propres	**0.7**	**-5.1**
Noms communs	**15.3**	**+13.7**
Adjectifs	**3.8**	**+16.2**
Adj. participe passé	*0.6*	*+142.7*
Pronoms	**18.3**	**-7.0**
Pronoms personnels	*9.4*	*-12.3*
Déterminants	**14.2**	**+12.8**
Articles	*10.0*	*+11.1*
Nombres	*1.8*	*+10.7*
Possessifs	*0.7*	*+4.1*
Démonstratifs	*0.5*	*+41.8*
Indéfinis	*1.2*	*+28.4*
Adverbes	**9.2**	**-23.5**
Prépositions	**12.9**	**+22.8**
Conjonctions	**5.6**	**-22.5**
Mots étrangers	**0.1**	**+19.6**

Tous les écarts sont statistiquement significatifs… Comme il y a quelque 70 locuteurs différents dans le corpus de la négociation raisonnée et plus de 300 dans celui du français oral, les excédents et les déficits ne peuvent provenir de caractéristiques individuelles propres à certains enquêtés.

La première surprise provient du net déficit en adverbes et en conjonctions. Par exemple, là où les “Français de France” utilisent 100 adverbes, les enquêtés québécois n'en mobilisent que 76.5 : près d'un quart en moins ! Parmi les principaux adverbes, seule la construction “ne…pas” (ou “ne… plus”, “ne… que”, etc.) résiste à peu près à cette érosion. Pour le reste, voici la liste des principaux adverbes significativement sous-employés par les enquêtés québécois par rapport au corpus “hexagonal”. Le classement est fait par ordre de “spécificité” décroissante (cf. plus bas) :

> bon, enfin, puis, bien, alors, non, pas, oui, maintenant, même, tout, trop, forcément, si, cher, là-bas, ici, dehors, justement, peu, jamais, mieux, petit, surtout, très, moins, certainement, partout, quelquefois, franchement, apparemment, combien, dessus, déjà, plus, pourtant, grand, pratiquement, là-haut, simplement, pourquoi, automatiquement, parfois, complètement, uniquement, heureusement, demi, effectivement, voire, au-dessus, encore, spécialement, mal, d'abord, normalement, dedans, aujourd'hui, vraiment, vachement, par-là, toujours, a priori, rarement, notamment, largement, malheureusement, pis, bientôt, ailleurs, comment, presque, obligatoirement, hyper, longtemps, systématiquement, hier, plutôt, souvent, directement, tôt, facilement, bas, suffisamment, énormément, actuellement…

En fait, la plupart des adverbes — de temps, lieu ou manière — entretiennent des liens de substitution avec le groupe nominal (voir Arrivé, 1986). Au lieu de “maintenant”, on dira : “à l'heure présente”, ou “de manière certaine” au lieu de “certainement”, etc.

Comme on peut s'y attendre, le tableau ci-dessus suggère que le déficit en adverbes se trouve compensé par l'excédent des adjectifs… Cependant, cette substitution n'est pas sans conséquence du point de vue de la communication : les propos ont un aspect plus accompli et moins tendu dont nous donnerons plus bas quelques exemples ;

— le déficit en conjonctions signale une faible coordination des propos (Antoine, 1958-1962). Parmi les principales conjonctions, seuls “comme” et “lorsque” échappent à ce déficit considérable. Outre “que”, voici la liste des principales conjonctions très significativement sous-employés : *donc, et, mais, quand, puisque, parce que, sinon, si, car, ou, soit*…

Ces deux déficits majeurs proviennent-ils des enquêtés ? Par exemple, pour la coordination, les interviewés présentent-ils les choses “à plat”, sans trop se soucier de les relier entre elles ? Ou bien s'agit-il d'une caractéristique propre au “français du Québec” qui le différencierait fondamentalement du “français de France” ?

L'excédent considérable des participes passés s'explique très probablement par les conditions particulières de l'enquête. La quasi-totalité des entretiens ont eu lieu après la fin de la négociation et la question essentielle était la suivante :

“Pouvez-vous me raconter de façon assez détaillé ce qui s'est passé [à partir du début de la négociation et] jusqu'à maintenant ?”

En revanche, l'excédent en participes présents et en infinitifs est une caractéristique propre à la plupart des enquêtés québécois. Ces formes verbales sont celles qui se rapprochent le plus du groupe nominal. Elles sont à mettre en corrélation avec l'excédent considérable en adjectifs issus du participe passé : négociation (ou méthode, approche) *raisonnée*, moment *donné*, formation *continue*, etc. Par rapport au verbe fléchi, ces formes verbales “dégradées” présentent l'avantage d'effacer, totalement ou partiellement, l'action et l'agent de celle-ci. La tension s'en trouve diminuée. Le propos tend vers l'accompli.

Le déficit important en pronoms personnels peut être rattaché à ce même phénomène de relative “dépersonnalisation” des propos.

Comme on le voit, la simple comparaison des catégories grammaticales a permis de soulever un problème passionnant ! Si l'absence de corpus “nationaux” représentatifs ne permet pas de conclure en faveur de l’hypothèse régionale, celle-ci se trouve plutôt confirmée par la comparaison des vocabulaires.

III. Comparaison des vocabulaires

Pour comparer le vocabulaire des enquêtés avec celui de la population de référence, plusieurs procédés sont envisageables.

Le plus évident consiste à comparer la fréquence d'emploi de chaque mot dans les deux corpus en appliquant les tests statistiques classiques pour la comparaison des fréquences d'un caractère dans deux populations différentes : khi2, loi normale pour les vocables les plus fréquents (fréquence supérieure à 30), loi de Poisson pour les autres, etc. Outre le recours, toujours malaisé, à des tables, on remarquera que ces instruments ne donnent

que des approximations. Ce sont des palliatifs inventés à une époque où l'absence d'ordinateur interdisait que l'on puisse envisager le calcul direct.

La loi normale semble fournir un cadre intellectuel logique : on considère les entretiens comme des échantillons extraits aléatoirement d'une urne de Bernouilli constituée par le corpus de référence. Le prélèvement de l'échantillon n'affecte pas le contenu de l'urne : le tirage d'un mot est un événement indépendant de ceux qui l'ont précédé ou qui le suivent… L'espérance mathématique d'un événement — par exemple : probabilité qu'un mot *X* se trouve *n* fois dans l'échantillon — et la déviation standard autour de cette valeur centrale sont aisément calculables. Mais cela suppose que la taille des échantillons prélevés soit très petite par rapport à la dimension de l'urne (afin que le prélèvement n'affecte pas son contenu). La dimension des corpus disponibles rend une telle hypothèse intenable. Par exemple, la taille du “français oral” est de 2,4 millions de mots et celle de l'enquête “négociation raisonnée” de 410 000 mots. De plus, tout texte en langue naturelle comporte une proportion considérable de mots de faible fréquence : aussi grand que soit le corpus de référence, il contiendra toujours une majorité de vocables apparaissant une fois ou très rarement. Le tirage d'un de ces mots “rares” aura une influence évidente sur le contenu de l'urne et sur les épreuves suivantes.

Il est donc nécessaire d'utiliser la loi hypergéométrique — ou “tirage sans remise” (le tirage d'un vocable modifie son espérance mathématique de figurer dans les tirages suivants) — et d'inclure explicitement le corpus sous revue dans l'urne. Ce calcul est inspiré de celui proposé par P. Lafon pour les “spécificités du vocabulaire” (Lafon, 1984 et Labbé et Labbé, 1994).

Soit :
— le corpus de référence (A) composé de N_a occurrences (taille en “mots”) ;
— le sous-corpus étudié (B avec B ∈ A) composé de N_b occurrences ;
— un vocable i quelconque de fréquence absolue F_{ia} dans A et F_{ib} dans B.

Si les mêmes lois de composition sont en œuvre dans la population générale (A) et dans la sous-population étudiée (B), alors le vocable i aura une fréquence théorique dans B — ou “espérance mathématique” (E_{ib}) — qui sera fonction de sa fréquence dans A pondérée par le rapport entre la taille de B et celle de A :

$$E_{ib(u)} = F_{ia} * U \quad avec \quad U = \frac{N_b}{N_a}$$

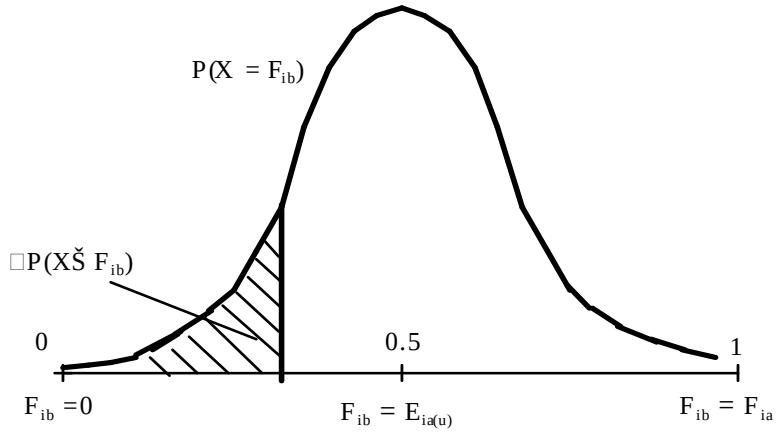
Si la fréquence constatée (F_{ib}) est différente de la fréquence attendue (E_{ib}), quand peut-on dire que le vocable est significativement sur-employé ou sous-employé dans B par rapport à A ? Pour répondre à cette question, il faut considérer la probabilité de l'événement observé F_{ib} par rapport à l'événement attendu (E_{ib}). Cette probabilité suit une loi hypergéométrique de paramètres F_{ia}, F_{ib}, N_a et N_b:

$$(1)\quad P(X = F_{ib}) = \frac{\begin{matrix}\lceil F_{ia} \rceil \\ \lfloor F_{ib} \rfloor\end{matrix} \begin{matrix}\lceil N_a - F_{ia} \rceil \\ \lfloor N_b - F_{ib} \rfloor\end{matrix}}{\begin{matrix}\lceil N_a \rceil \\ \lfloor N_b \rfloor\end{matrix}}$$

F_{ib} peut varier entre 0 — aucune occurrence du vocable dans B — et F_{ia} : toutes les occurrences du vocable sont observées dans le sous-corpus ($0 \leq F_{ib} \leq F_{ia}$).

En développant (1), on constate que le calcul n'a de sens que si $F_{ia} < N_b$ et $F_{ia} < (N_a - N_b)$. La seconde borne va de soi (l'urne doit être nettement plus grande que le corpus sous revue). La première borne signifie que le calcul doit porter sur de grands corpus, ou que, si B est petit, on doit exclure du calcul les vocables les plus fréquents (les “mots-outils”).

A condition que N_a, F_{ia} et N_b soient suffisamment grands, les valeurs de cette probabilité se distribueront selon la fameuse courbe en cloche, avec un mode pour $F_{ib} = E_{ia(u)}$ (graphique ci-dessous).



En arrêtant le cumul des valeurs de P lorsque $X = F_{ib}$, la probabilité pour que le vocable i suive la même loi de fréquence dans le sous-corpus B et dans le corpus de référence A sera la surface comprise sous la courbe (S). Celle-ci varie entre zéro (le vocable n'est pas attesté dans B) et 1 (toutes les occurrences du vocable i se manifestent dans B).

Un vocable sera significativement sur-employé dans B lorsque S aura une valeur supérieure à .975 ou à .995 suivant que l'on choisira un risque d'erreur de 5% ou de 1%. La liaison entre B et le vocable i sera d'autant plus forte que S sera plus proche de 1. A l'inverse, une valeur inférieure à .005 (ou à .025 si l'on choisit un risque d'erreur de 5%) signifiera que le vocable i est significativement sous-employé dans le corpus sous revue.

Avec cette méthode, le vocabulaire spécifique au corpus “négociation raisonnée” apparaît clairement. Voici par exemple, les substantifs et les adjectifs les plus caractéristiques de ce corpus (avec une chance d'erreur inférieure à 1 sur 1000) :

> *Substantifs* : négociation, problème, gens, chose, temps, partie, syndicat, comité, solution, travail, façon, convention, formation, moment, niveau, personne, intérêt, table, point, employeur, employé, confiance…
>
> *Adjectifs* : raisonné, syndical, bon, patronal, collectif, traditionnel, donné, nouveau, capable, autre, différent, important, difficile, clair, général, long, facile, partiel, intéressant, prêt, évident, conjoint…

La négociation raisonnée (entre les parties syndicales et patronales) est donc le premier thème ; le second tourne autour des “problèmes” (avoir…, poser…, régler…).

En effet, la même méthode peut être appliquée aux syntagmes répétés (Pibarot & Labbé, 1998). Elle fait apparaître le troisième thème principal de ces entretiens, thème qui tourne autour de la “confiance”. Il s'agit notamment de la confiance dans le fait que l'autre partie respectera les règles du jeu, mais aussi la confiance des mandataires —

l'employeur d'un côté, les membres du syndicat de l'autre — envers les représentants et, surtout, la confiance en soi-même, dans la justesse et la solidité de ses positions…

Cette même étude appliquée aux groupes verbaux fait apparaître un déficit très net en “modalisateurs” — les verbes pseudo-auxiliaires suivis d'un infinitif (sur le modèle “pouvoir faire, aller voir…”) — sauf, justement, pour “aller” et “pouvoir”. Les principales spécificités négatives sont : “savoir”, “croire”, “falloir”, “vouloir”, “devoir”. L'abondance relative en constructions “verbe + verbe” est généralement la marque d'un discours tendu. Autrement dit, sauf pour les modalités orientées vers l'action et le possible, les Québécois semblent nettement moins “tendus” que les Français…

Cette conclusion peut sembler paradoxale. En Amérique du Nord, la négociation collective passe pour très conflictuelle. En effet, le droit du travail n'a pas du tout l'extension qu'il a en France : l'essentiel des droits des salariés sont définis par les contrats collectifs. Les enjeux de la négociation sont donc importants, on discute souvent très longtemps, de manière animée ; la plupart des grèves surviennent à cette occasion. Alors ? la faible tension du discours serait-elle le résultat des méthodes raisonnées qui seraient parvenues à “détendre l'atmosphère” dans des rencontres pourtant traditionnellement conflictuelles ? Ou plus probablement, serait-ce la traduction de traits culturels profondément différents des deux côtés de l'Atlantique ? Seule la constitution de corpus représentatifs permettra de résoudre cette intéressante énigme ainsi que beaucoup d'autres, beaucoup plus fondamentales, concernant notre langue.

Reférences

Antoine G. (1958 et 1962). *La coordination en français*. Paris. d'Atrey (deux tomes).

Arrivé M., Gadet F., Galmiche M. (1986). *La grammaire d'aujourd'hui. Guide alphabétique de linguistique française*. Paris. Flammarion.

Berger G., Leselbaum N. dir. (2002*). La prévention des toxicomanies en milieu scolaire : éléments pour une évaluation*. Montpellier. CNDP.

Bergeron J.-G. et Bourque R. (1996). “ L’impact de la formation sur les pratiques de la négociation raisonnée ” in Bélanger J. et Al, *Innover pour gérer les conflits*. Sainte-Foy. Presses de l’université Laval.

Bergeron J.-G. et Bourque R. (1998). “ La formation et la pratique de la négociation collective raisonnée au Québec : esquisse d’un bilan ” in Deschênes et Al, *Négociation en relation du travail. Nouvelles approches*. Sainte-Foy. Presses du l’Université du Québec.

Bergeron J.-G., Bourque R. et White F. (2003). “ Empirical Assessment of an Interest-based Bargaining Training Program in Labor-Management Relations ”. A paraître dans *Relations industrielles*.

Bergeron J.-G., Labbé D. (2000). “L'évaluation de la négociation raisonnée par les acteurs. Une analyse lexicométrique” (XVI^{e} Congrès de l'AISLF, Québec, juillet 2000). Reproduit dans Bernier C. et Al. *Formation, relations professionnelles à l'heure de la société-monde*. Paris-Québec. L'Harmattan - Les Presses de l'Université Laval. 2002 : 239-252.

Blanche-Benveniste C. et Jeanjean C. (1987). *Le français parlé. Transcription et édition*. Paris, Didier.

Burnard L. (1995). *Users Reference Guide for the British National Corpus*. Oxford. Oxford University Computing Service.

Crowdy S. (1993). “Spoken Corpus Design”, *Literary and Linguistic Computing*, 8-4 : 259-266.

Demazière D. et Dubar C. (1997). *Analyser les entretiens biographiques. L’exemple des récits d’insertion*. Paris. Nathan.

Douglas F. M. (2003). “The Scottish Corpus of Texts and Speech : Problems of Corpus Design”, *Literary and Linguistic Computing*, 18 (1-2) : 23-37.

Habert B., Fabre C., Issac F. (1998). *De l'écrit au numérique*. Paris. Masson.

Jenny J. (1997). “Méthodes et pratiques formalisées d'analyse de contenu et de discours dans la recherche sociologique française contemporaine. Etat des lieux et essai de classification”, *Bulletin de Méthodologie Sociologique*, 54 (mars 1997) : 64-122.

Kucera K. (2002). “The Czech National Corpus : Principles, Design and Results”, *Literary and Linguistic Computing* (17-2) : 245-258.

Lafon P. (1984). *Dépouillements et statistiques en lexicométrie*. Genève-Paris. Slatkine-Champion.

Labbé D. (1990). *Normes de saisie et de dépouillement des textes politiques*. Grenoble. Cahier du CERAT.

Labbé C. et Labbé D. (1994). *Que mesure la spécificité du vocabulaire ?*. Grenoble, CERAT. Repris dans *Lexicometrica*, 3, 2001.

Labbé D. (2001). “Normalisation et lemmatisation d'une question ouverte. Les femmes face au changement familial”, *Journal de la Société Française de Statistique*, 142-4 : 37-57.

Labbé D. (2002). *Analyse des représentations du confort électrique à partir d'un corpus d'entretiens* (Rapport pour le GREST-EDF). Grenoble. CERAT.

Meurman-Solin A. (2001). “Structured text corpora in the study of language variation and change”, *Literary and Linguistic Computing* (16) : 5-27.

Nelson G. (1997). “Standardizing Wordforms in a Spoken Corpus”, *Literary and Linguistic Computing* (12-2) : 79-85.

Pibarot A. et Labbé D. (1998). “Les syntagmes répétés dans l'analyse des commentaires libres”, in Mellet S. (ed*), 4e Journées d'analyse des données textuelles*. Nice. Université de Nice : 507-516.

Pionchon S. (2001). *Les Françaises et la politique* (Thèse pour le doctorat de science politique). Institut d'Etude Politique. Grenoble.

Annexe

Le corpus “ français oral ”

(voir en bibliographie les ouvrages correspondants)

Les Français(es) et la politique (Pionchon) :
32 entretiens : 345 752 mots, 6 540 vocables différents

La négociation raisonnée au Québec (Bergeron & Bourque, 1996, 1998, 2003) :
61 entretiens : 409 225 mots, 6 591 vocables différents

La prévention des toxicomanies en milieu scolaire (Berger & Leselbaum) :
15 entretiens : 92 992 mots, 4 255 vocables différents

Confort électrique réalisé par les sociologues du Grets-EDF en six enquêtes (Labbé, 2002) :
201 entretiens : 1 270 307 mots, 10 904 vocables différents

Questions ouvertes dans un sondage auprès des femmes divorcées réalisé par l'INED (Labbé,2001) :
3000 enquêtés : 56 107 mots, 2 786 vocables différents

Questions ouvertes dans un sondage auprès des citoyens belges sur la droite et la gauche :
1000 enquêtés : 22 294 mots , 1 706 vocables différents

Divers :
6 entretiens : 115 494 mots, 4 922 vocables différents

Total transcriptions de l'oral :
322 entretiens et deux sondages : *2 264 498* mots, 16 809 vocables différents.