La Langue Française dans le Dictionnaire Bilingue : Méthodes d'Evaluation

Margareta Kastberg Sjöblom
ILF-CNRS
Bases, Corpus et Langage (UMR 6039)
UFR Lettres, Arts et Sciences Humaines
98, bd E. Herriot. B.P. 209
06204 Nice Cedex 3
kastberg@unice.fr

Résumé
L'établissement de la nomenclature et des exemples qui figurent dans le dictionnaire est en grande partie fondé sur la conscience linguistique du lexicographe ainsi que sur la connaissance qu'il manifeste du public auquel il s'adresse. Le choix des mots et des phrases est inévitablement assez subjectif et reflète la propre conception du monde de l'auteur et sa perception personnelle de la réalité. Par conséquent, les deux langues figurant dans le dictionnaire bilingue sont toujours marquées par les concepteurs de l'ouvrage. Comment saisir le français diffusé dans un dictionnaire bilingue ? Comment définir ce portrait de la langue française, cette vitrine vers l'étranger, qu'il constitue indiscutablement ? Nous nous proposons ici d'essayer de définir le discours français d'un dictionnaire français-suédois étudié avec l'aide des méthodes lexicométriques. Le logiciel Hyperbase autorise en effet un ensemble de traitements linguistiques sur des corpus de textes prédéfinis. Le corpus phraséologique d'un dictionnaire, c'est-à-dire les phrases et les exemples à l'intérieur du dictionnaire, est une forme de corpus clos et pourrait même être considéré comme un genre de discours qui s'adapte efficacement à ce type d'analyse et qui permet de prendre en considération simultanément - et de façon impartiale - la totalité du corpus.

1. Le Dictionnaire et sa Nomenclature

Bien que la fonction première du dictionnaire bilingue ne soit pas de définir la norme linguistique, il reste néanmoins une institution sociale de caractère normatif. Il autorise des mots, des constructions, des phrases, des sens, et inversement il en condamne ou en écarte d'autres. L'établissement de la nomenclature et des exemples qui figurent dans le dictionnaire bilingue - qui est toujours plus ou moins restreint - est, en grande partie, fondé sur la conscience linguistique du lexicographe ou des collaborateurs du groupe rédactionnel, ainsi que sur la connaissance qu'il manifeste du public auquel il s'adresse. Le lexicographe décrit la langue selon sa conception scientifique, ses théories linguistiques, et le grand public ne remet pas en question cette "institution".

En effet, tout texte reflète son auteur et le dictionnaire ne fait pas exception à cette vérité. Le choix des mots et des phrases est inévitablement assez subjectif et trahit peu ou prou la conception du monde des auteurs et leur perception de la réalité à un moment donné de l'histoire. Les conséquences de ce postulat sont multiples : il s'avère souvent que le lexicographe donne une importance "démesurée" à certaines catégories lexicales au détriment d'autres. Le vocabulaire d'un champ sémantique spécifique peut faire défaut dans

un dictionnaire bilingue donné, tandis que d'autres champs sémantiques peuvent être richement représentés dans la phraséologie interne du même ouvrage.
Ces distorsions peuvent mettre en question l'efficacité opérationnelle du dictionnaire car l'usager a souvent du mal à trouver les lexèmes qui correspondent à ses besoins d'expression dans la vie actuelle, et la présence hypertrophiée d'un champ sémantique spécifique peut contribuer à donner un teint à un dictionnaire et refléter ainsi une image conventionnelle, voire stéréotypée de la langue française éloignée de la réalité moderne avec, par exemple, une abondance de termes de galanterie, d'élégance, etc. (cf. M. Kastberg Sjöblom ; 2003).
Il convient peut-être - à une époque où le français diminue nettement en tant que langue de communication internationale - de s'interroger sur l'image de la langue française que nous diffusons à l'étranger, notamment par le biais de nos dictionnaires. Les phrases à l'intérieur du dictionnaire forment un ensemble de discours qu'il serait utile d'évaluer.
Cependant une vision globale de ce corpus difficilement saisissable qu'est le discours disparate du dictionnaire bilingue demande le recours à une technique qui dépasse l'œil humain. Comment saisir le français dans un dictionnaire bilingue ? Comment définir ce portrait de la langue française qu'il dessine ? Nous nous proposons ici d'essayer de définir le discours français d'un dictionnaire avec l'aide des méthodes lexicométriques et de la technique de linguistique de corpus.

2. La Technique Lexicométrique

Les techniques de la lexicométrie, et au-delà celle de la logométrie, qui se sont inspirées en grande partie des méthodes de la statistique lexicale développées dans les années 1960 notamment par Pierre Guiraud et de Charles Muller (Guiraud 1954 ; Muller 1968), offrent aujourd'hui de nouvelles perspectives à toute étude de corpus, quel que soit son genre. Des travaux pionniers dans ce champ de recherche, qui s'intéressaient surtout à des genres littéraires comme le théâtre classique, s'est développée une technique de travail aujourd'hui, non seulement qui s'applique à divers corpus littéraires, mais sert aussi de plate-forme technique pour des études sur l'évolution de la langue française, pour des analyses génériques et sous-génériques ainsi que pour des études sur le langage politique, journalistique, etc.
La constitution de grands corpus textuels, leur structuration en bases de données exploitables pour le traitement quantitatif ainsi que pour diverses recherches linguistiques et stylistiques ont été une des préoccupations les plus importantes dans ce domaine ces dernières années. La plupart des corpus sont désormais étiquetés morpho-syntaxiquement afin d'élargir les possibilités - après la lexicométrie - vers une analyse plus ample de différents aspects de la langue.
En outre, le travail de développement et d'amélioration d'outils logiciels a fait franchir un seuil qualitatif considérable au traitement lexicométrique. Ce progrès qualitatif balaie ainsi un certain nombre d'objections quant à la légitimité de traiter seulement la surface des textes, et offre pour la première fois un outil de traitement statistique complet (des chaînes de caractères aux isotopies sémantiques) du discours.
Le logiciel Hyperbase, conçu et développé par Étienne Brunet (2001), autorise un ensemble de traitements sur des corpus de textes prédéfinis. Le corpus phraséologique d'un dictionnaire, c'est-à-dire les phrases et les exemples à l'intérieur du dictionnaire, est une

forme de corpus clos et pourrait même être considéré comme un genre de discours qui s'adapterait, après un traitement informatique adéquat, à ce genre d'analyse qui permet de prendre en considération simultanément la totalité du corpus.
Le traitement informatique permet l'exploitation documentaire ainsi que l'obtention de contextes et de concordances. La distribution d'un mot (ou d'un groupe de mots) 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, notamment sur la richesse lexicale, l'étude des hapax, la distance (ou connexion) lexicale, la corrélation chronologique et les spécificités internes et externes. La comparaison externe s'appuie ici sur une partie sélectionnée du corpus du *Trésor de la langue française* (*T.L.F.*). Une fonction thématique recense tous les termes situés dans l'environnement immédiat d'un mot donné et en compare la fréquence réelle avec celle que laissent attendre les modèles de distribution aléatoire, fournissant ainsi une liste des mots en excédent probablement attirés par le mot-pôle. L'analyse factorielle des listes et du dictionnaire permet, grâce à la représentation graphique et "géographique", une vision synthétique des multiples accords ou oppositions qui lient les mots entre eux.
Ces différentes analyses permettent de comparer, au niveau exogène, différents dictionnaires unilingues et bilingues, différentes époques etc. et, au niveau endogène, de donner une vision de l'unité et de l'homogénéité des exemples, de leur longueur, de leur diversité ou, au contraire, des thèmes récurrents qu'ils véhiculent.
Certes, la nomenclature d'un dictionnaire constitue bien – comme nous l'avons dit dans l'introduction – une forme de discours qui mérite d'être étudiée sous ces différents aspects qui permettront un point de vue plus différencié et impartial sur la langue que le dictionnaire bilingue diffuse. Nous nous intéresserons ici, au titre d'exemple d'application possible, aux phrases extraites du dernier grand dictionnaire français-suédois paru sur le marché suédois, *Norstedts stora svensk-franska ordbok* (1998)[1].

3. Le Dictionnaire Français-Suédois

Il est aisé de constater qu'un dictionnaire, même le plus complet, ne contient jamais tous les vocables de la langue. Le temps de le rédiger, de nouveaux mots ont fait leur apparition. Les dictionnaires "traînent" des vocables dont personne ne se sert plus et qui se transmettent de génération en génération comme autant d'éléments morts. La part de ce décalage est cependant plus ou moins grande selon les dictionnaires.
Dans le cas des dictionnaires français-suédois, ce décalage a été marquant et lorsque nous avons élaboré *Norstedts stora svensk-franska ordbok*, nous sentions tous intuitivement que le "parc dictionnairique " était non seulement suranné, mais qu'il était également très teinté par ses auteurs et leur culture, reflétant un monde trop marqué par la haute bourgeoisie qui diffusaient l'image d'une France idéalisée ; ce qui nous a incités à un effort de modernisation considérable (cf. Kastberg Sjöblom, 2003).
L'informatique a radicalement changé le travail de lexicologie et de lexicographie. L'informatisation des inventaires rend désormais possible ce qui semblait chimérique il y a seulement vingt ans. Les logiciels dont se servent les maisons d'édition pour l'élaboration du dictionnaire offrent désormais de nombreuses possibilités telles que l'indexation des vedettes, le contextage automatique des mots, le tri ultrarapide de millions d'occurrences,

etc. et offrent par là-même un accès facile aux corpus dictionnairiques, jusqu'à présent très peu exploités dans les études linguistiques quantitatives.

Ainsi, nous avons exploité le corpus informatisé de l'inventaire français du dictionnaire *Norstedts stora svensk-franska ordbok* qui est aujourd'hui l'ouvrage de référence dans la paire de langues français/suédois. Ce corpus, après un travail d'adaptation à notre logiciel Hyperbase (version 5.5), a été soumis à un traitement lexicométrique "traditionnel".

4. Données Statistiques

4.1. Structure du corpus et distribution des occurrences.

Notre corpus est constitué par les phrases et les syntagmes qui constituent la partie française des articles du dictionnaire *Norstedts stora svensk-franska ordbok* ; il englobe 159.263 occurrences[2] reparties sur les 26 lettres de l'alphabet, choisies ici comme les jalons des différents sous-corpus. Cette répartition, qui s'aligne sur le principe traditionnel dictionnairique, permet non seulement la vérification de l'importance donnée à chaque lettre de l'alphabet dans l'œuvre, mais aussi la comparaison proportionnelle avec d'autres dictionnaires bilingues ou unilingues. Pour obtenir une première vue générale de la structure quantitative de notre dictionnaire nous observons la distribution relative de chaque sous-corpus (c'est-à-dire l'inventaire correspondant à chaque lettre de l'alphabet) qui est la suivante :

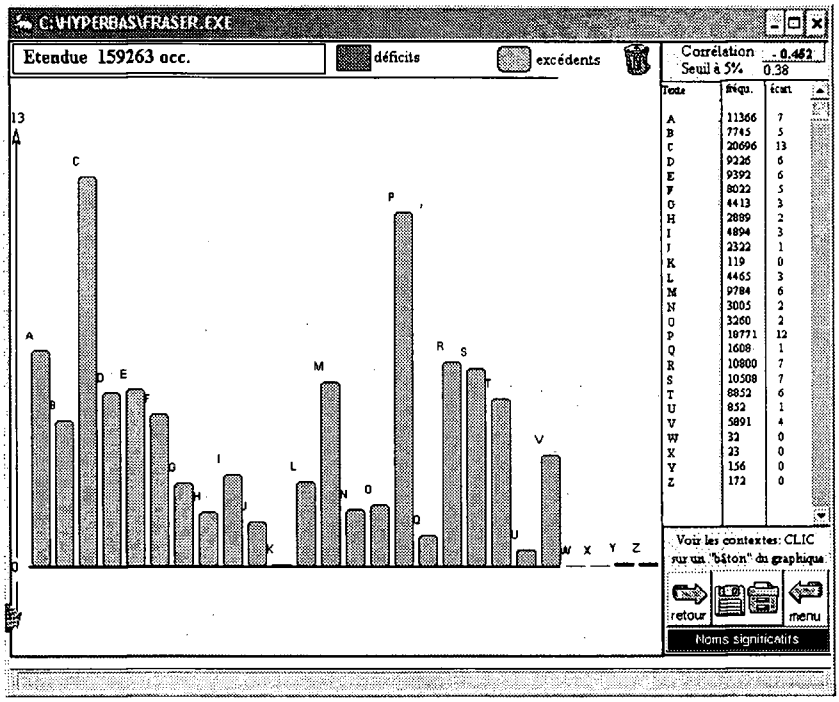

Figure 1: Etendue du corpus

La proportion de chaque lettre donnée dans un dictionnaire dépend évidemment des caractéristiques de la langue. Ces résultats n'ont rien d'étonnant, sachant par exemple que la lettre *c* et la lettre *p* occupent de nombreuses pages dans tout dictionnaire français[3]. En revanche, dans une perspective comparative cette analyse peut se révéler très intéressante car elle permet de façon efficace de constater et prévenir d'éventuels problèmes de distorsion dans les inventaires dictionnairiques des différents dictionnaires.
L'analyse de la distribution des hautes fréquences, c'est-à-dire les mots les plus employés du corpus[4], permet aussi de distinguer les caractéristiques d'un discours.
Parmi les hautes fréquences dans notre dictionnaire, *Norstedts stora svensk-franska ordbok*, nous trouvons les mots grammaticaux que l'on trouve dans tout corpus : *de*, *à*, *la*, *un*, *le*, *en*, *se*, *les*, *une*, *des*, etc. Il est toutefois notable que parmi les 100 mots les plus fréquents, on ne relève pas un seul substantif ou adjectif comme ceux que l'on trouve en tête de liste dans n'importe quel corpus littéraire, journalistique ou politique.
Nous nous trouvons en effet, dans le dictionnaire, dans un espace essentiellement verbal et la riche fréquence des pronoms témoigne également de cette réalité[5]. Des verbes comme *faire*, *être*, *avoir*, *prendre*, *mettre* et *donner* en tête de liste reflètent non seulement la description (par les verbes d'état) mais aussi une activité incessante à l'intérieur des articles dictionnairiques. *On fait*, *on prend*, *on met* et *on donne* dans des exemples qui privilégient nettement les pronoms à la première et à la deuxième personne.
Toutefois, il va de soi qu'une fréquence ne saurait devenir "caractéristique" que comparée à une fréquence théorique, donc par référence à un texte ou à un corpus plus étendu que celui qui est sous analyse ; cet ensemble plus grand est alors pris comme source du modèle théorique.

4.2. Spécificités du vocabulaire

Pour connaître, de façon impartiale, les mots spécifiques et caractéristiques de notre corpus, nous utiliserons comme point de comparaison dans cette étude *Frantext* (qui s'appuie sur les fréquences du *Trésor de la langue française* avec ses 86 millions d'occurrences), et plus précisément le corpus du XX^{ème} siècle. Dans Hyperbase, en effet, ces données sont insérées en tant que norme et servent de base de calcul en indiquant la différence entre deux grandeurs, celle des fréquences dans notre corpus dictionnairique et celle de *Frantext*. Les valeurs obtenues mettent en relief les excédents et les déficits du vocabulaire phrastique français du dictionnaire par rapport à celui de *Frantext*[6]. Il convient de rappeler à ce sujet que la comparaison avec l'usage observé dans le *Trésor de la langue française* doit être interprétée prudemment. D'une part, le *T.L.F.* reflète l'usage littéraire de la langue, dans un registre relativement élevé, et d'autre part toutes les formes n'ont pas été soumises à la comparaison, parce que le calcul de l'écart réduit perd de sa légitimité quand la fréquence théorique est trop faible, ce qui dépend certes de la taille du corpus traité, mais aussi de la fréquence du mot en question. Prenant ces considérations en compte, le traitement informatique nous permet d'extraire le vocabulaire spécifique positif et négatif de notre corpus, c'est-à-dire les mots particulièrement suremployés et sous-employés par rapport à la base de référence, afin de nous donner une idée précise des thèmes traités et non traités.

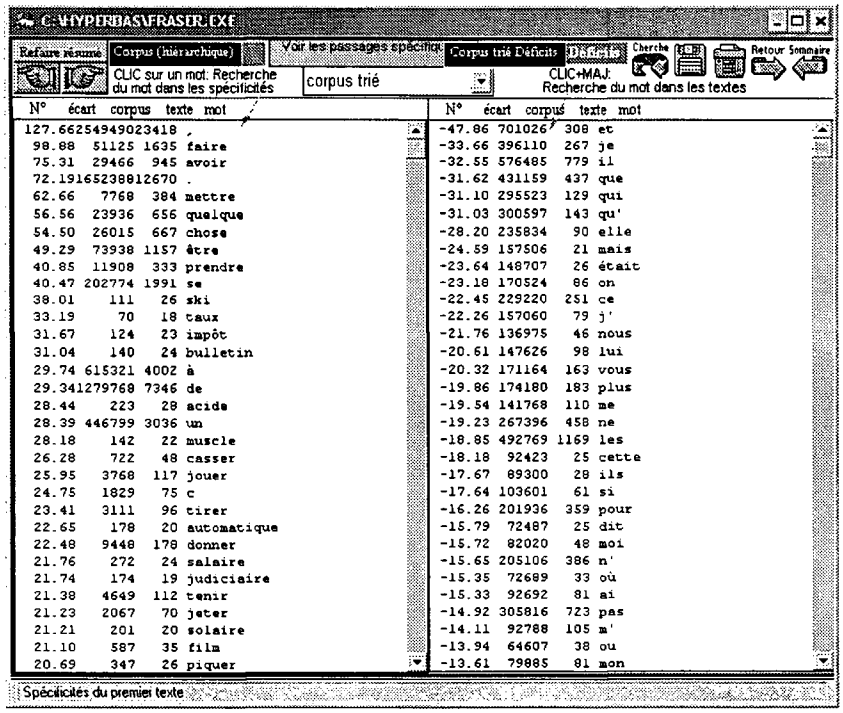

Figure 2 : Le vocabulaire spécifique du corpus.

Les mots qui se trouvent en tête de liste des spécificités sont toujours les verbes d'action (*faire*, *mettre*, *prendre*), ce qui n'étonnera aucun lexicographe qui s'applique à donner des exemples utiles et pratiques. Mais il s'agit en réalité aussi d'un phénomène caractéristique du français ; la nominalisation – beaucoup moins fréquente dans la langue suédoise – qui oblige le lexicographe à fabriquer des exemples afin de traduire des substantifs comme *affront*, *appel* etc. avec des constructions verbales : *faire un affront*, *faire appel* etc. à verbe support.

Lorsqu'il s'agit des substantifs et des adjectifs, les résultats sont assez étonnants et reflètent une réalité qui est peut-être moins celle de la France idéalisée à laquelle les dictionnaires antérieurs nous avaient habitués, que celle de la Suède elle-même ! Avec *ski*, *taux*, *impôt* etc. en haut de la liste nous sommes en effet non seulement dans la réalité climatique nordique mais également dans la réalité fiscale pesant sur le peuple le plus taxé du monde. De ce point de vue on peut dire sans doute que l'objectif de l'équipe rédactionnelle de *Norstedts fransk-svenska Ordbok* qui voulait moderniser l'image du français véhiculée dans le dictionnaire bilingue a été atteint ; mais d'autres distorsions, sans doute plus légères, n'ont pas été évitées.

En revanche, que nous trouvions les pronoms personnels parmi les mots les plus déficitaires (la colonne de droite) lors d'une comparaison exogène n'a rien d'étonnant, compte tenu des nombreux exemples impersonnels qui caractérisent tout dictionnaire, ce que reflète aussi

l'importance relative de verbes à l'infinitif. Ces deux phénomènes sont en effet très liés, étant donné l'absence du pronom dans des constructions à l'infinitif comme *se laisser faire, faire démarrer un moteur*, etc. qui offrent des définitions détachées de tout ancrage énonciatif et non actualisées. Ceci suggère au passage que, dans ce dictionnaire du moins, les définitions pèsent plus lourd que les exemples (qui, eux, reproduisent des énoncés actualisés). C'est probablement une des différences structurelles fondamentales entre un dictionnaire bilingue et un dictionnaire monolingue, surtout de grande ampleur comme le *T.L.F.*

5. Contextes et concordances

Revenons un instant à nos spécificités positives, c'est-à-dire de fréquence excédentaire par rapport à *Frantext*. Pourquoi est-il donné une telle importance au mot *bulletin*, qui figure parmi les mots les plus spécifiques du corpus ? (cf. figure 2.). La fonction de recherche en contexte du logiciel permet un recensement immédiat de son emploi, qui dépasse en effet l'emploi à l'intérieur de l'article dont il est la vedette :

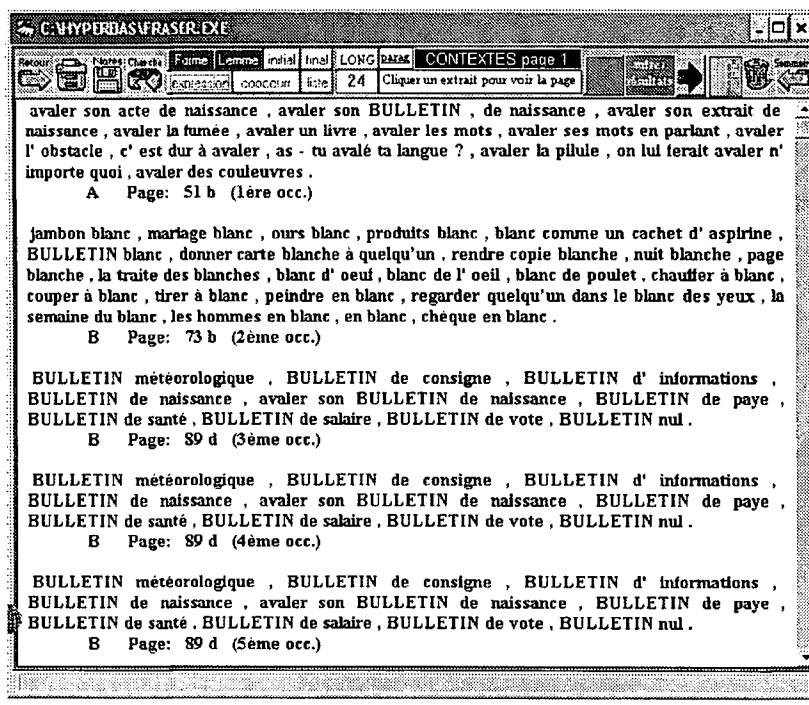

Figure 3 : Quelques contextes du mot *bulletin*.

La largeur du contexte permet aisément de deviner l'insertion de la première occurrence dans l'article *avaler* et celle de la deuxième dans l'article *blanc*. En cherchant les concordances d'une occurrence, la liste que l'on obtient au bout du traitement par le logiciel

permet aussi de situer facilement la présence du mot à l'intérieur des articles autres que celui qui comporte le mot en question en tant que vedette, grâce la colonne de gauche qui indique le sous-corpus (ici la lettre de l'alphabet en question et le numéro de page). Le tableau ci-dessous, prenant comme exemple le mot *impôt* permet de repérer facilement la présence de ce mot dans d'autres articles que celui consacré à ce mot :

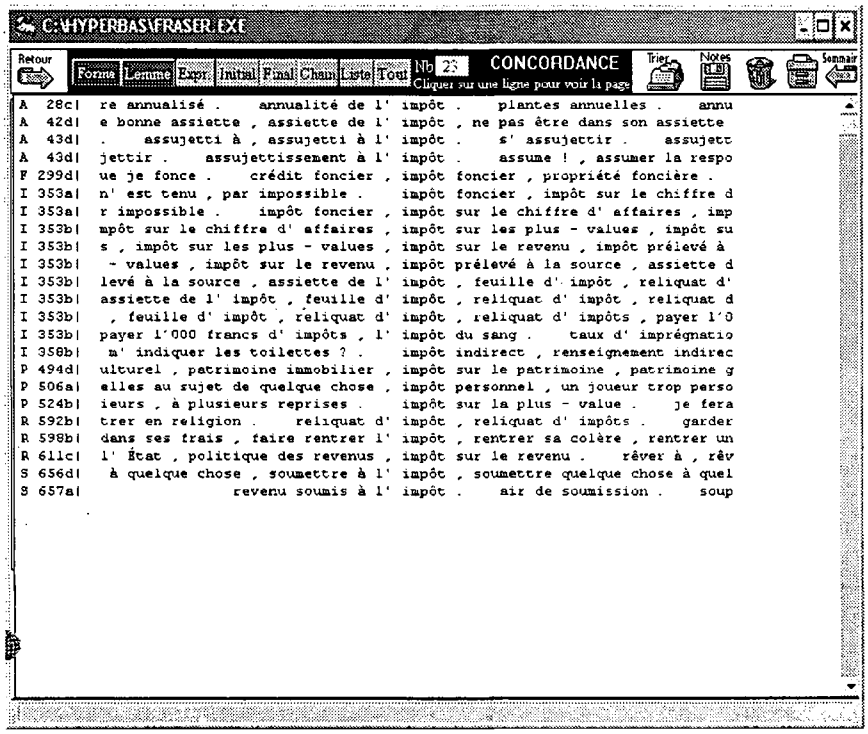

Figure 4 : Concordance du mot *impôt*.

Nous voyons ici aisément que des syntagmes comme *annualité de l'impôt*, *assiette de l'impôt*, *assujetti à l'impôt*, *assujettissement à l'impôt* etc. sont répertoriés sous les vedettes respectives ; *annualité*, *assiette*, *assujetti*, *assujettissement* etc., sans pour autant être présents dans l'article consacré à la vedette *impôt*, tandis qu'un syntagme comme *impôt foncier* ou *reliquat d'impôt* figurent sous les deux vedettes qui les composent. La fonction "trier" permet d'ailleurs d'afficher clairement ces syntagmes répété au sein du dictionnaire, en regroupant les occurrences du mot cherché selon l'ordre alphabétique du mot qui le précède ou du mot qui le suit.

En tant que lexicographes nous nous posons souvent la question de savoir sous quelle vedette répertorier un syntagme ou une phrase. Faut-il les faire figurer sous les vedettes respectives de tous leurs constituants ? Ou bien sous une seule, et dans ce cas laquelle ? L'application du logiciel ne permet guère de résoudre ce problème délicat, mais il fournit au

chercheur une manière exacte, fiable et immédiate de recenser les différentes compositions lexicales.
A ce propos, une autre difficulté dans l'élaboration des dictionnaires dans la paire français-suédois est celle des unités nominales et de la position du complément. Le complément de nom en suédois est antéposé et attaché au nom, formant ainsi une seule lexie. En français, en revanche, les compléments de détermination, les compléments de l'adjectif et les compléments du nom se trouvent détachés des mots qu'ils déterminent, précédés par une préposition qui leur donne un statut différent de celui des correspondantes unités nominales suédoises, ce qui crée de grandes difficultés voire un déséquilibre notable, aux niveaux de la nomenclature et de la bidirectionnalité du dictionnaire.
On trouve donc d'emblée les mots et les vedettes du dictionnaire répertoriés selon des systèmes différents dans le dictionnaire français-suédois et dans le dictionnaire suédois-français. Il s'agit ici non seulement d'un contraste dans la structure formelle des deux langues.

5. Conclusion

Le corpus du dictionnaire bilingue est un espace interculturel, et sa fonction est de constituer un pont entre deux peuples, deux cultures. Cet outil de communication par excellence mérite ainsi d'être étudié de façon efficace et systématique. L'approche lexicométrique et le recours aux méthodes quantitatives peuvent en effet constituer un complément intéressant dans la recherche lexicographique.
Les recherches en méthodologie lexicométrique, le développement de logiciels de traitement textuel ainsi que la constitution de bases de données par le recueil de textes, mis en chantier depuis plusieurs décennies, arrivent progressivement à maturité. Ils peuvent donc maintenant pleinement remplir leur mission : servir d'outil à la recherche linguistique dans toutes ses dimensions.
En effet, les progrès apportés aux logiciels ouvrent aujourd'hui la voie à de nouvelles perspectives de recherche jusqu'alors peu exploitables de façon systématique ; ils offrent désormais, et pour la première fois, un outil de traitement statistique complet du discours. Les exemples ici n'en sont qu'une première ébauche. Nous aimerions encore élargir ce projet avec l'exploitation systématique du discours dictionnairique, notamment des dictionnaires bilingues. Celle-ci permettrait une analyse objective des documents existants, tant en termes quantitatifs, qu'en termes qualitatifs. Cette analyse pourrait à la fois déboucher sur une évaluation impartiale dans un domaine qui n'est pas toujours exempt d'influences idéologiques, et sur l'élaboration d'outils susceptibles d'aider les lexicographes dans la rédaction des dictionnaires à venir. Enfin la constitution et l'implémentation de bases de données dictionnairiques constitue une sauvegarde patrimoniale notamment dans le cas de grands dictionnaires de référence.

Remerciements

Je tiens à remercier la maison d'édition Norstedts à Stockholm qui m'a aimablement mis à disposition le corpus de travail. Mes remerciements vont aussi à Étienne Brunet qui m'a fourni une aide technique précieuse dans l'adaptation des données pour le logiciel Hyperbase.

Notes

1. *Norstedts stora fransk-svenska ordbok*, le Grand Dictionnaire français-suédois (1998) Stockholm, Norstedts, (74.000 mots et phrases selon l'éditeur).

2. La définition de *mot* en linguistique est ambiguë et n'a pas de délimitation satisfaisante. Selon la terminologie de Ch. Muller (1977 ; 4), les *mots* sont les "unités dont la suite constitue un énoncé ou un texte ; il s'agit essentiellement d'une unité graphique séparée des unités voisines par un blanc ou un signe de ponctuation." Dans notre étude, nous employons, comme le propose Étienne Brunet (1978 ; 23), le terme d'*occurrence* ; l'ensemble des occurrences d'un texte est symbolisé par *N*.

3. *Le* Petit Robert (1972) :
a = 132,5 pages, b = 70,5 p., c = 192,25 p., d = 124,5 p., e = 145 p., f = 89 p., g = 58 p.,
h = 43,75 p., i = 76,75 p., j = 18,5 p., k = 3 p., l = 54 p., m = 117,75 p., n = 34,75 p., o = 43,5 p.,
p = 214,5 p., q = 11,25 p., r = 145,5 p., s = 146,25 p., t = 119,25 p., u = 10,75 p., v = 61,75 p.,
w = 1,5 p., x = 1 p., y = 1 p., z = 3,5 p.

4. Dans les études statistiques, pour effectuer des analyses quantitatives différentes, les fréquences absolues ne suffisent pas : il est important de connaître l'étendue de son corpus et de ses parties. En effet, les valeurs de N (occurrences) et de V (vocables) ne sont pas liées par une relation fixe. Or, les calculs effectués par le logiciel Hyperbase permettent de mesurer l'étendue des textes dans le corpus en prenant en compte ces contraintes. Les calculs du poids relatif, c'est-à-dire l'espérance mathématique de l'événement : occurrence d'un mot dans le texte considéré (P) et non-occurrence de ce mot dans le même texte (Q=1-P), permettent l'emploi des lois classiques de la lexicométrie, principalement la loi normale et la loi binomiale (Brunet, 2001), et elles servent aux calculs de pondération dans les différents traitements statistiques de notre étude.

5. La corrélation dans toute analyse lexicométrique entre le verbe et le pronom est par ailleurs bien documentée (cf. M. Kastberg Sjöblom 2002 ; 339-341).

6. La méthode consiste, pour un fragment d'un texte, à calculer l'écart réduit de chacun des vocables du texte par rapport à la sous-fréquence théorique, et à classer ceux-ci en fonction de cet écart réduit. On obtiendra ainsi, en tête de liste, le vocabulaire caractéristique positif du fragment, c'est-à-dire l'ensemble des vocables dont la sous-fréquence est plus élevée que la fréquence dans le texte ne le fait prévoir ; et en fin de liste, le vocabulaire caractéristique négatif (cf. Ch. Muller 1968 : 204).

Références

Adam J.-M. 1992. *Les textes : Types et prototypes*, Paris, Nathan, coll. « fac. linguistique ».

Béjoint H., Thoiron Ph. 1996. *Les dictionnaires bilingues*, Louvain-la-Neuve, Aupelf-Uref, Editions Duculot. Aupelf-Uref, Editions Duculot.

Biber D., Conrad S., Reppen R. 1998. *Corpus linguistics, Investigating Language, Structure and Use*, Cambridge, Cambridge Approaches to Linguistics.

Brunet É. 1981. *Le vocabulaire français de 1789 à nos jours*, Paris - Genève, Champion - Slatkine.

Brunet É. 2001. *Hyperbase*, Manuel de référence, version 5.0, Nice, CNRS-INaLF, "Bases, corpus et langage" (UMR 6039).

Guiraud P. 1954. Les caractères statistiques du vocabulaire, Paris, puf.

Hammar Th. 1943. *Svensk-fransk ordbok, Dictionnaire suédois-français*, Stockholm, Svenska Bokförlaget P.A. Norstedt & Söner.

Kastberg Sjöblom M. 2002. *L'écriture de J.M.G. Le Clézio, une approche lexicométrique*, Nice, Université de Nice-Sophia Antipolis.

Kastberg Sjöblom M. 2002. 'Le choix de la lemmatisation. Différentes méthodes appliquées à un même corpus', in A. Morin et P. Sébillot (eds.), *JADT 2000, 6èmes Journées internationales d'Analyse statistique des Données Textuelles*, Saint-Malo, Irisa, Inria, p. 391-402.

Kastberg Sjöblom M. 2003. 'Les dictionnaires dans la paire français–suédois ; une approche culturelle' in A.-M. Laurian et T. Szende (eds.) *Dictionnaires bilingues et interculturalité*, Editions Peter Lang, Collection « Etudes contrastives », Berne, en cours de publication.

Malrieu D. et Rastier F. 2002. 'Genres et variations morphosyntaxiques', in A. Martin Municio (ed.), *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, Logroño, Fundación San Millán de la Cogolla,

Muller Ch. 1967. *Le vocabulaire du théâtre de Pierre Corneille, Étude de statistique lexicale*, Paris (réédition Genève, Slatkine Reprints, 1993).

Muller Ch. 1977. *Principes et méthodes de statistique lexicale*, Paris, Hachette.

Muller Ch.1979. 'Calcul des probabilités et calcul d'un vocabulaire' in Muller Ch. (ed.), *Langue française et linguistique quantitative*, Genève, Slatkine.

***Norstedts stora svenk-franska och fransk-svenska ordbok**, le Grand Dictionnaire français-suédois et suédois-français*. 1998. Stockholm, Norstedts.

***Prismas stora svensk-franska och fransk-svenska ordbok**, le Grand Dictionnaire français-suédois et suédois-français*. 1999. Stockholm, Prisma.

Rastier F. 1991. *Sémantique et recherches cognitives*, Paris, puf, coll. Formes sémiotiques.

Svensén B. 1987. *Handbok i lexikografi, Principer och metoder i ordboksarbetet*, Stockholm, Tekniska Nomenklaturcentralen, Esselte Studium.

Vising J. 1950. *Svensk-fransk ordbok, Dictionnaire français-suédois* Stockholm, Svenska Bokförlaget Albert Bonnier.