Ce que l'analyse lexicométrique d'un dictionnaire français-suédois nous enseigne sur le discours dictionnairique

Margareta Kastberg Sjöblom

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

Abstract

The establishment of the nomenclature and the examples which appear in the dictionary is mainly founded on the linguistic awareness of the lexicographer and on his or her knowledge of the public to whom it is addressed. The choices of the words and the sentences are inevitably rather subjective and reflect very often the authors personal perception of reality. The systematic study of dictionary corpora reveal large linguistic as well as cultural differences inbetween different dictionaries. The purpose of this paper is to define and study the French language of a corpus from a French-Swedish dictionary *Norstedts Stora Fransk-svenska Ordbok*, by exploring lexicostatistical methods. The exploration and development of statistical methods can in fact be very useful in dictionary compilation, particularly in the study of dictionaries where it opens new avenues of research for a larger audience in the field of lexicography. The phraseological content of a dictionary, i.e. the sentences and the examples in the context of different words in the dictionary, is indeed a form of closed corpus and could even be regarded as a specific discourse or genre, which, after an adequate data-processing treatment, adapts perfectly to corpus linguistics. The quantitative method, which makes it possible to take into account the totality of the corpus simultaneously, gives a synthetic and impartial view of the language communicated by the dictionary.

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 de la réalité. Par conséquent, les deux langues figurant dans le dictionnaire bilingue sont toujours marquées par les concepteurs de l'ouvrage. Comment saisir ces nuances ? Et comment définir 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 dictionnairique français à partir d'un dictionnaire français-suédois étudié avec l'aide des méthodes lexicométriques. 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.

Mots-clés : lexicométrie, genres de discours, logométrie, lexicographie, phraséologie, dictionnaires bilingues

1. Le Dictionnaire et sa Nomenclature

Les dictionnaires sont généralement associés par le consultant à deux réflexes. Le premier consiste à évoquer "le" dictionnaire comme s'il s'agissait d'un ouvrage unique, or contrairement aux idées reçues et aux manières de dire, un dictionnaire n'est pas un objet unique et total, il n'existe pas de dictionnaire unique, mais DES dictionnaires avec les visées lexicographiques qui sont propres à chacun. Le second réflexe courant est celui de "la vérité

absolue", avec le postulat tacite suivant : tout ce qui est mentionné dans les dictionnaires est indiscutable, et le mot qui n'y figure pas n'existe pas. C'est évidemment oublier un peu trop rapidement que les auteurs de dictionnaires enregistrent des nomenclatures très différentes en fonction de la taille de l'ouvrage, du nombre de volumes, et des choix qu'ils se fixent quant au regard sur les mots, descriptif ou normatif par exemple.

Bien que la fonction première du dictionnaire bilingue ne soit pas de définir la norme linguistique, il reste néanmoins, tout autant que le dictionnaire monolingue, une institution sociale de caractère normatif. En effet, 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'ils manifestent du public auquel ils s'adressent.

Le choix des mots et des exemples 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. (M. Kastberg Sjöblom, 2003).

La nomenclature d'un dictionnaire constitue bien 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 celui que le dictionnaire bilingue diffuse. 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 de définir le discours français extrait de la phraséologie d'un dictionnaire avec l'aide des méthodes lexicométriques.

Le corpus phraséologique d'un dictionnaire, c'est-à-dire les phrases et les exemples à l'intérieur du dictionnaire, est certes 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.

Différentes analyses quantitatives permettent, en effet, 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.

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].

2. 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 puisque le lexique d'une langue est une liste ouverte. Le temps de le rédiger, de nouveaux mots ont fait leur apparition. En revanche, 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*, les rédacteurs sentaient 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 diffusait l'image d'une France idéalisée ; ce qui nous a incité 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.

C'est pourquoi 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 au logiciel Hyperbase (version 5.5), a été soumis à un traitement lexicométrique "traditionnel".

3. Données Statistiques

3.1. 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

[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, 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." (1977 : 4). Dans notre étude, nous employons, comme le propose É. Brunet, le terme d'occurrence ; l'ensemble des occurrences d'un texte est symbolisé par N.

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 :

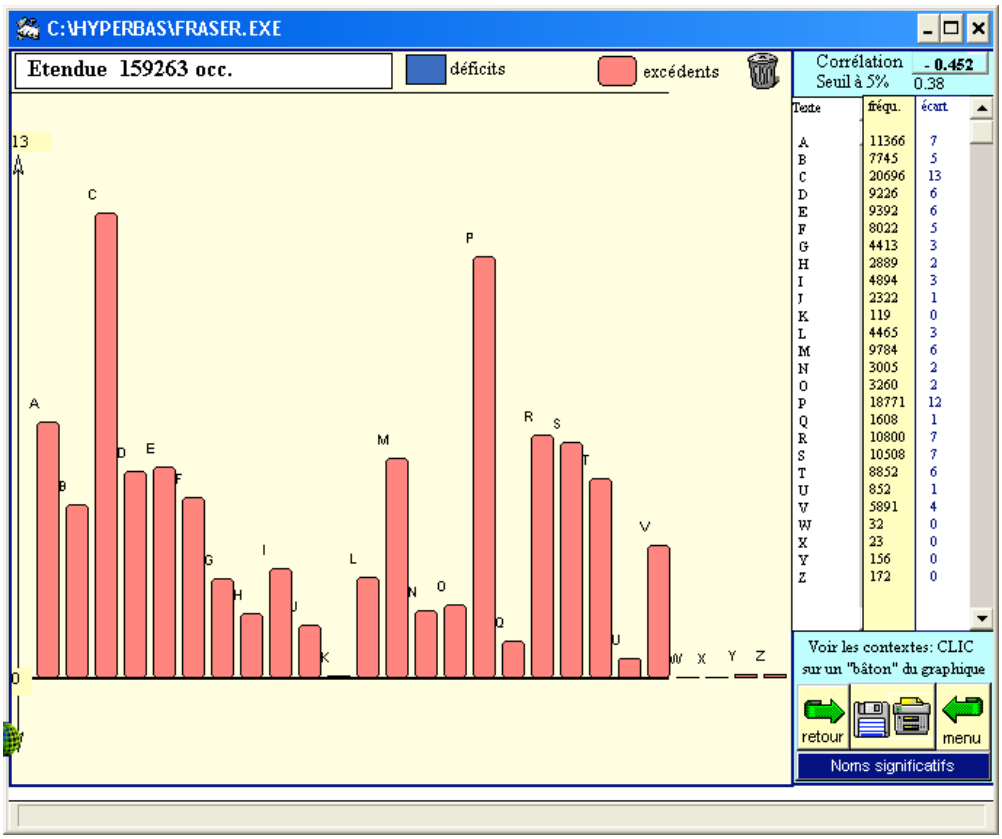

Figure 1: Étendue du corpus

La proportion de chaque lettre donnée dans un dictionnaire dépend évidemment des caractéristiques de la langue. Les résultats observés 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.

[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

Parmi les hautes fréquences constatées dans notre dictionnaire, 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 pourtant 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.

3.2. 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 afin de nous donner une idée précise des thèmes traités et non traités.

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 (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).

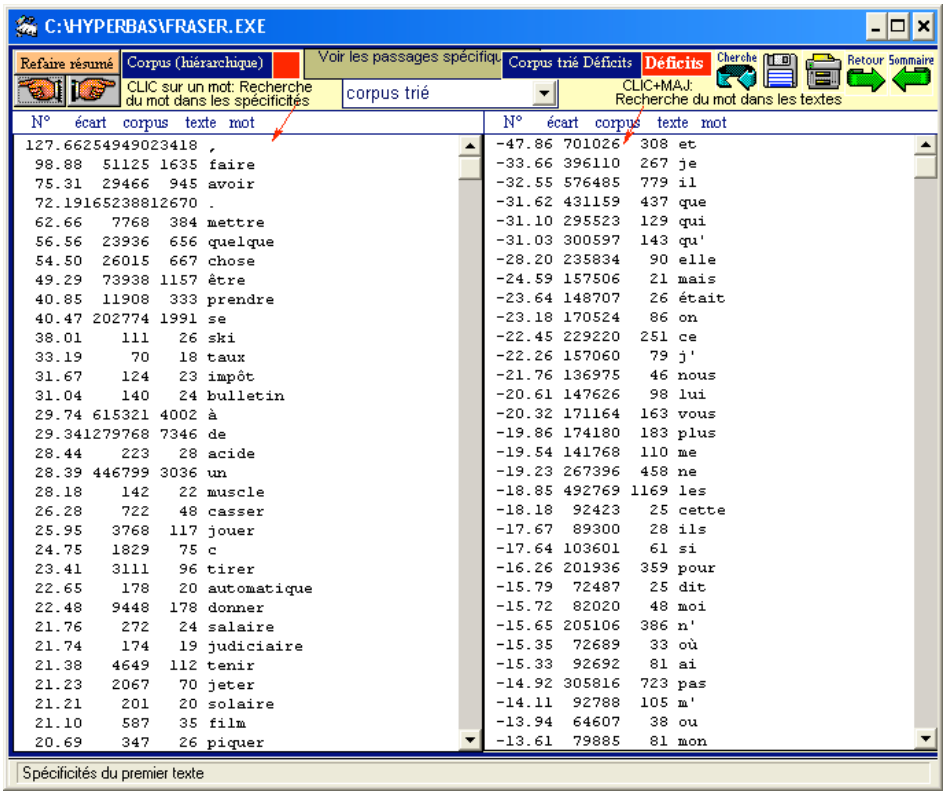

Figure 2 : Le vocabulaire spécifique du corpus

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 surprenants 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.*

4. Contextes, concordances et collocations

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 :

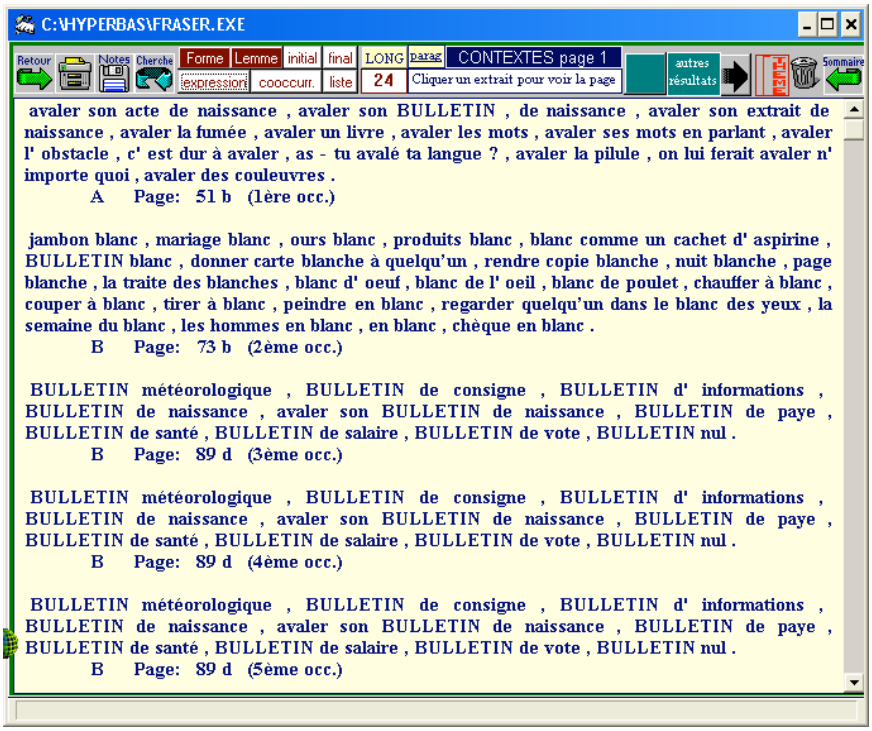

Figure 3 : Quelques contextes du mot bulletin

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 à 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 qui y est directement consacré.

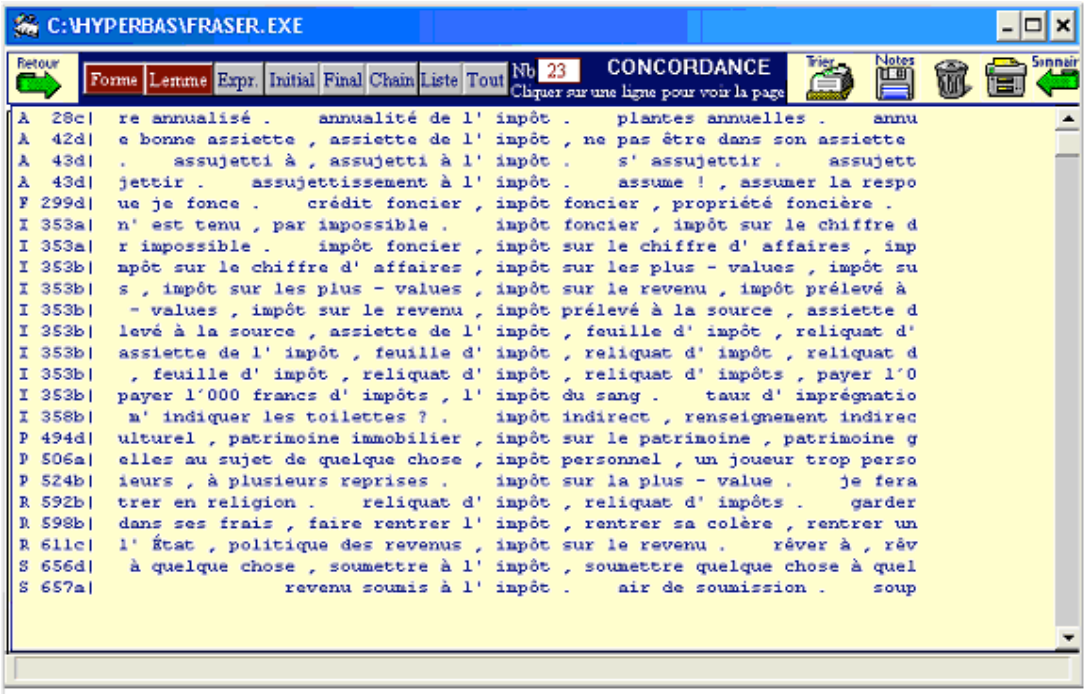

Figure 4 : Concordance du mot impôt

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.

À ce propos, une autre difficulté de l'élaboration des dictionnaires dans la paire français-suédois est celle des unités nominales et de la position du complément. Il s'agit ici d'un contraste dans la structure formelle des deux langues. 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 unités nominales suédoises correspondantes, 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. Prenons l'exemple *blouson* : sous la vedette française nous trouvons les *blouson*

noir, *blouson d'aviateur* et *blouson de cuir*. Dans le dictionnaire suédois-français, l'exemple *svart jacka - blouson noir* ne figure pas et les autres exemples sont répertoriés en tant que vedettes dans la nomenclature : *flygarjacka* et *läderjacka*. L'inventaire par l'outil lexicométrique permet de recenser les entrées lexicales suivantes : *bäddjacka* ("liseuse"), *dunjacka* ("doudoune"), *jeanssjacka* ("blouson en jean") *regnjacka* ("blouson imperméable"), *skidjacka* ("blouson de ski"), *sportjacka* ("blouson de sport") et *vinterjacka* ("blouson d'hiver"). Nous voyons ici que ce que l'on considère en français comme des exemples d'une vedette devient en suédois des entrées lexicales à part entière. Le contenu phrastique à l'intérieur des paragraphes est donc d'autant plus important dans le dictionnaire partant du système d'une langue romane comme le français, où des objets ou des concepts bien définis en tant que vedettes dans les dictionnaires ayant une langue germanique en tant que langue source sont traités comme des exemples d'utilisation.

La notion de collocation devient aussi en lexicographie bilingue extrêmement importante. L'imprévisibilité d'une collocation est un facteur important, non seulement à l'intérieur d'un système langagier, mais d'autant plus dans la dimension bilingue. Pour celui qui doit s'exprimer dans une langue étrangère, l'imprévisibilité de la collocation constitue une grande difficulté, étant donné qu'il est impossible de savoir si une collocation dans la langue maternelle peut se traduire mot à mot vers la langue cible. Par exemple, l'équivalent français de la collocation suédoise *hålla ett föredrag* ("tenir une conférence") est formé avec un autre verbe : *faire une conférence*. S'ajoute à ceci une difficulté supplémentaire, celle que la collocation dans une langue ne correspond pas forcement à une collocation dans l'autre langue où qu'il peut s'agir d'une combinaison libre et banale.

Pouvoir employer correctement des collocations dans une autre langue que la sienne est un signe de connaissance approfondie de celle-ci, d'autant qu'un emploi erroné donne une impression d'incertitude. Le rôle du dictionnaire est ici primordial. Il s'agit de répertorier et de fournir au consultant des collocations utiles dans les différents articles du dictionnaire. Les outils lexicométriques peuvent ici être extrêmement utiles et donner une aide précieuse à l'utilisateur du dictionnaire.

La définition de collocation est assez floue, et les collocations sont de nature très différente. Certaines collocations sont considérées comme transparentes (Grossmann F., Tutin A., : 2003), c'est-à-dire qu'elles comporte des collocatifs facilement compréhensibles mais imprédictibles du point de vue lexical et/ou syntaxique, comme dans les expressions *avoir faim*, *prendre peur*. Déjà Bally dans son *Traité de stylistique française* (1909) cite les exemples suivants : *grièvement blessé*, *? gravement blessé*, *gravement malade*, *? grièvement malade*, où les adverbes *gravement* et *grièvement* ont à peu près le même sens sans pour autant être interchangeables et ils sont encore présentés comme des prototypes de collocations.

Les recherches de segments répétés, d'expressions ou de cooccurrences qu'offrent les logiciels *Lexico3* et *Hyperbase* permettent d'extraire de façon systématique et extrêmement rapide les collocations répertoriées dans le dictionnaire. La recherche des quatre combinaisons d'occurrences entre *gravement*, *grièvement*, *malade* et *blessé* par *Hyperbase* permet immédiatement de constater une formulation maladroite dans notre dictionnaire. Les collocations *grièvement blessé*, *gravement malade* et *grièvement malade* sont absentes du corpus, tandis la présence *gravement blessé* est tout à fait discutable :

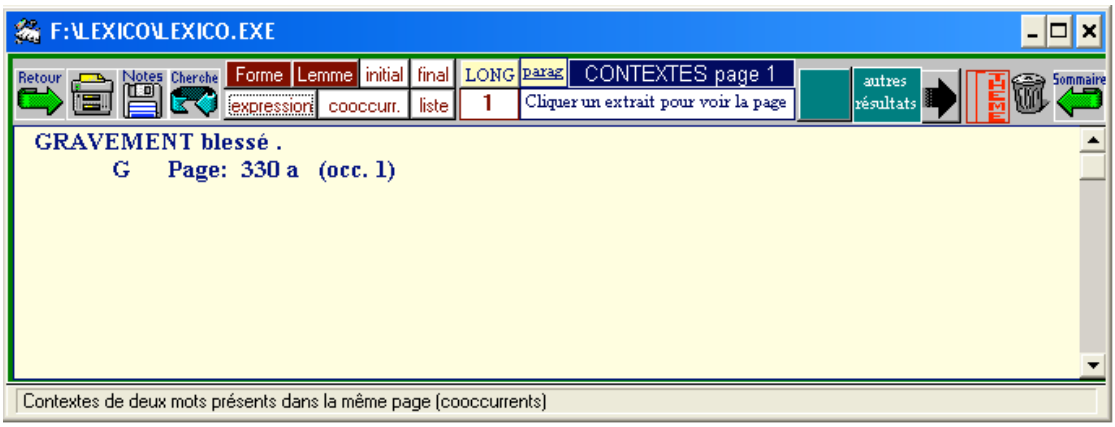

Figure 5 : Cooccurrence des mots gravement *et* blessé

La langue contient aussi des collocations opaques, qui comprennent des collocatifs imprédictibles et démotivés sémantiquement, comme dans *peur bleue, nuit blanche, colère noire*, etc. : la collocation est difficilement décodable et tout à fait imprédictible. La recherche automatique de cooccurrences d'Hyperbase permet ici de repérer les deux collocatifs *peur* et *bleu* dans le corpus dictionnairique et nous pouvons aisément constater que les lexicographes ont répertorié la collocation deux fois dans le dictionnaire, sous la vedette *bleu* et sous la vedette *peur*.

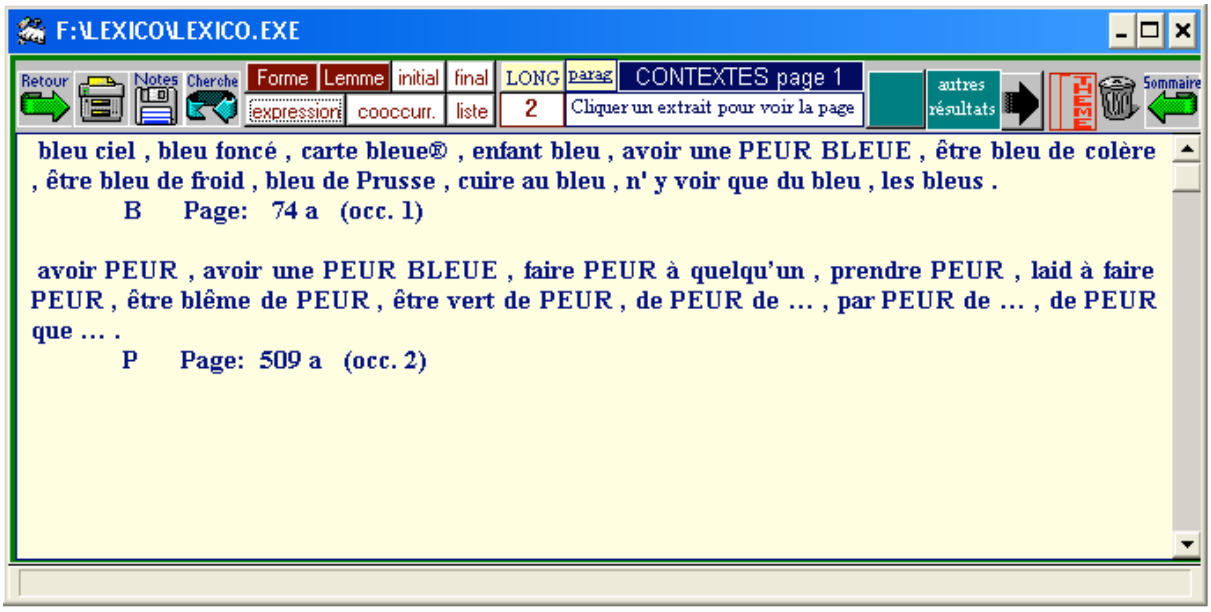

Figure 6 : Cooccurrence des mots peur *et* bleu

La mise en contexte de la collocation permet aussi de cerner les figements syntaxiques caractéristiques de certaines collocations. En effet, les propriétés syntaxiques des collocations sont très variables, et d'une manière générale, elles ne semblent pas permettre de circonscrire aussi nettement ces associations que les propriétés sémantiques. Certaines collocations présentent un figement syntaxique important, c'est le cas de *peur bleue* où seule la constellation *une peur bleue* semble possible, des combinaisons telles que **la peur est bleue*

et **une peur très bleue* n'étant guère admises. L'outil lexicométrique nous permet ici de vérifier le travail du lexicographe et la technique lexicographique. Dans le cas de *peur bleue,* le figement de la collocation est mis en évidence par la construction d'un syntagme : *avoir une peur bleue*. En revanche, dans le cas des combinaisons comme *enfant bleu* ou *carte bleue* la construction syntagmatique n'a pas été jugée comme utile.

Outre de prospecter dans le corpus dictionnairique, la recherche assistée par ordinateur des collocations dans des grands corpus de référence permet au lexicographe de vérifier ou d'obtenir une attestation d'un emploi de manière efficace. Les techniques de la lexicométrie et les méthodes de linguistique de corpus constituent ainsi une excellente aide pour le lexicographe dans son travail de rédaction ainsi que pour l'étude comparative dictionnairique.

5. Conclusion et perspectives

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. 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 et dictionnairique.

En effet, les progrès apportés aux logiciels ouvrent aujourd'hui la voie à de nouvelles perspectives de recherche en lexicographie jusqu'alors peu exploitées de façon systématique. Les exemples ici n'en sont qu'une première ébauche. Nous aimerions – malgré les difficultés considérables d'obtention de ces corpus à valeur commerciale –élargir encore 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 déboucher à la fois 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 constituent une sauvegarde patrimoniale, notamment dans le cas de grands dictionnaires de référence.

Références

Atkins S., Zampolli A. (éds.). (1994). *Computional Approches to the Lexicon*. Oxford, Oxford University Press.

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

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

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

Grossmann F., Tutin A. (éds.). (2003). *Collocations – analyse et traitement*. Amsterdam, Éditions De Werelt, Travaux et Recherche en linguistique appliquée, Série E – N°1.

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

Kastberg Sjöblom M. (2003). « Les dictionnaires dans la paire français–suédois ; une approche culturelle ». In Laurian A.-M. et Szende T. (eds.) *Dictionnaires bilingues et interculturalité*. Éditions Peter Lang, Collection « Études contrastives », Berne, en cours de publication.

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

Rastier F. (1991). *Sémantique et recherches cognitives*. Paris, PUF, coll. Formes sémiotiques.

Salem A. (1987). *Pratique des segments répétés*. Paris, Publications de l'INaLF, collection "St.Cloud", Klincksieck.

Szende Thomas (éd.). (2003). *Les écarts culturels dans les dictionnaires bilingues*. Paris, Honoré Champion.

Svensén B. (2004). *Handbok i lexikografi, Ordböcker och ordboksarebte i teori ochpraktik*. Stockholm, Norstedts Akademiska förlag.

Remerciements : Je tiens à remercier la maison d'édition Norstedts à Stockholm qui m'a aimablement mis à disposition le corpus de travail.